Reference Manual on Scientific Evidence

Fourth Edition

Volume I

FJC Version

Committee on Science for Judges—
Development of the Reference Manual
on Scientific Evidence, Fourth Edition

Committee on Science, Technology,
and Law

Policy and Global Affairs

FJC
Federal Judicial Center

NATIONAL ACADEMIES
Sciences
Engineering
Medicine

This project is funded in part by the Gordon and Betty Moore Foundation through Grant GBMF10224, and the National Science Foundation under Grant GBMF10224 to the National Academy of Sciences. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

The Federal Judicial Center contributed to this publication in furtherance of the Center's statutory mission to conduct research and continuing education programs for judicial branch employees for the improvement of judicial administration. This manual is provided as a resource to assist judges in understanding relevant scientific methods and principles that may arise in their cases. The views expressed are those of the authors and not necessarily those of the Federal Judicial Center and its Board.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine and Federal Judicial Center. 2025. *Reference Manual on Scientific Evidence: Fourth Edition* (FJC Version). Washington, DC.

Note: This is an edited version of the *Reference Manual on Scientific Evidence* (4th ed.) published on December 31, 2025. It is available on the Federal Judicial Center websites and is published in hard copy only for distribution to the federal judiciary and international judiciaries with which the Center works. Any other distribution or sale is a violation of copyright law.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. Tsu-Jae Liu is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

FEDERAL JUDICIAL CENTER

Congress established the Federal Judicial Center to support the efficient and effective administration of the federal courts through research and continuing education of judges and court employees. Its separate status within the judicial branch, its specific missions, and its specialized expertise enable it to pursue and encourage critical and careful examination of ways to improve judicial administration. The Federal Judicial Center has no policymaking or enforcement authority; its role is to provide accurate, objective information and education and to encourage thorough and candid analysis of policies, practices, and procedures.

By statute, the Chief Justice of the United States chairs the Center's Board, which also includes the director of the Administrative Office of the U.S. Courts and seven judges elected by the Judicial Conference of the United States. The Board appoints the Center's director and deputy director; the director appoints the Center's staff. Since its founding in 1967, the Center has had twelve directors. John S. Cooke was director from September 2018–August 2025; Judge Robin L. Rosenberg became director in August 2025. Clara Altman has been deputy director since 2018.

The organization of the Center reflects its primary statutory mandates. The Director's Office is responsible for the Center's overall management and its relations with other organizations. The Research Division examines and evaluates current and alternative federal court practices and policies. This research assists the Judicial Conference of the United States. The Center's research also contributes to its education mission. The Education Division plans and produces education and training for judges and court staff, including in-person programs, video programs, publications, and Web-based programs and resources. The Federal Judicial History Office helps courts and others study and preserve federal judicial history. The International Office provides information to judicial and legal officials from foreign countries and informs federal judicial personnel of developments in international law and other court systems that may affect their work. Two units of the Director's Office—the Information Technology Office and the Editorial & Information Services Office—support Center missions through technology, editorial, and design assistance, and organization and dissemination of Center resources.

For more information, see **www.fjc.gov**.

Committee on Science for Judges—Development of the Reference Manual on Scientific Evidence, Fourth Edition

NANCY D. FREUDENTHAL (*Co-Chair*), Judge, U.S. District Court for the District of Wyoming

FRED H. GAGE (NAS, NAM) (*Co-Chair*), Adler Professor, Laboratory of Genetics, The Salk Institute for Biological Studies

RUSS BIAGIO ALTMAN, Kenneth Fong Professor of Bioengineering, Genetics, Medicine, Biomedical Data Science, and (by courtesy) Computer Science, Stanford University

DAVID G. CAMPBELL, Senior Judge, U.S. District Court for the District of Arizona

ALICIA L. CARRIQUIRY (NAM), Distinguished Professor of Liberal Arts and Sciences and the President's Professor of Statistics, Iowa State University

LYNN R. GOLDMAN (NAM), Michael and Lori Milken Dean and Professor of Environmental and Occupational Health, Milken Institute School of Public Health, George Washington University

BRIAN W. KERNIGHAN (NAE), William O. Baker '39 Professor, Department of Computer Science, Princeton University

PRAMOD P. KHARGONEKAR, Vice Chancellor for Research and Distinguished Professor of Electrical Engineering and Computer Science, University of California, Irvine

GOODWIN LIU, Associate Justice, Supreme Court of California

SHOBITA PARTHASARATHY, Professor of Public Policy and Women's and Gender Studies, Ford School of Public Policy and Director, Science, Technology, and Public Policy Program, University of Michigan

PATTI B. SARIS, Judge, U.S. District Court for the District of Massachusetts

THOMAS D. SCHROEDER, Judge, U.S. District Court for the Middle District of North Carolina

DAVID S. TATEL, Judge (retired), U.S. Court of Appeals for the District of Columbia Circuit

National Academies Staff

ANNE-MARIE MAZZA, Project Director and Senior Director, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine

STEVEN KENDALL, Senior Program Officer, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine

RENEE DALY, Senior Program Assistant, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine (until April 2025)

DOMINIC LOBUGLIO, Senior Program Assistant, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine (until February 2023)

MATTHEW COWAN, Christine Mirzayan Science and Technology Policy Graduate Fellow (until May 2023)

Federal Judicial Center Staff

ELIZABETH C. WIGGINS, Director, Research Division, Federal Judicial Center

JASON A. CANTONE, Senior Research Associate, Federal Judicial Center

MEGHAN A. DUNN, Senior Research Associate, Federal Judicial Center

ELLEN URHEIM, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2025); Contractor with Federal Judicial Center (since September 2025)

REBEKAH PETROFF, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2024)

ALEXIS ALLEGRA, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2023)

RESHMINA WILLIAM, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2022)

Consultant

JOE S. CECIL (until June 2024)

Board of the Federal Judicial Center

THE CHIEF JUSTICE OF THE UNITED STATES (*Chair*)
KATHLEEN CARDONE, Judge, U.S. District Court for the Western District of Texas
R. GUY COLE, JR., Judge, U.S. Court of Appeals for the Sixth Circuit
SARA L. ELLIS, Judge, U.S. District Court for the Northern District of Illinois
RALPH R. ERICKSON, Judge, U.S. Court of Appeals for the Eighth Circuit
MICHELLE M. HARNER, Judge, U.S. Bankruptcy Court for the District of Maryland
SUZANNE MITCHELL, Magistrate Judge, U.S. District Court for the Western District of Oklahoma
LYNN WINMILL, Judge, U.S. District Court for the District of Idaho
ROBERT J. CONRAD, JR., Judge and Director of the Administrative Office of the U.S. Courts

Committee on Science, Technology, and Law

MARTHA MINOW (*Co-Chair*), Professor of Law, Harvard Law School
HAROLD VARMUS (NAS/NAM) (*Co-Chair*), Lewis Thomas University Professor, Meyer Cancer Center, Weill Cornell Medicine
DAVID APATOFF, Senior Partner, Arnold & Porter
ERWIN CHEMERINSKY, Dean and Jesse H. Choper Distinguished Professor of Law, University of California, Berkeley School of Law
ELLEN WRIGHT CLAYTON (NAM), Professor of Law, Professor of Health Policy, Vanderbilt University Medical Center
JOHN S. COOKE, Director, Federal Judicial Center (until August 2025)
JENNIFER EBERHARDT (NAS), Professor of Psychology, Stanford University
KENNETH C. FRAZIER, Chairman, Health Assurance Initiatives, General Catalyst
CAROL GREIDER (NAS/NAM), Distinguished Professor of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz
STEVE HYMAN (NAM), Harald McPike Professor of Stem Cell and Regenerative Biology, Core Institute Member and Director, Broad Institute of MIT and Harvard
JON M. KLEINBERG (NAS/NAE), Tisch University Professor, Cornell University
BARBARA McGAREY, Deputy General Counsel (retired), U.S. Department of Health and Human Services
ERNEST J. MONIZ, Cecil and Ida Green Professor of Physics and Engineering Systems, Post-Tenure, Massachusetts Institute of Technology
KIMANI PAUL-EMILE, Professor of Law, Fordham University Law School
K. SABEEL RAHMAN, Professor of Law, Cornell Law School
NATALIE RAM, Professor of Law, University of Maryland Francis King Carey School of Law
JULIE ROBINSON, Senior Judge, U.S. District Court for the District of Kansas
PATTI B. SARIS, Judge, U.S. District Court for the District of Massachusetts
VICKI SATO, Chair, VIRbio
BARBARA SCHAAL (NAS), Mary-Dell Chilton Distinguished Professor, Washington University in St. Louis
JOSHUA SHARFSTEIN, Vice Dean and Professor, Johns Hopkins Bloomberg School of Public Health
CLIFFORD TABIN (NAS), Leder Professor and Chair, Harvard Medical School

Staff

ANNE-MARIE MAZZA, Senior Director
STEVEN KENDALL, Senior Program Officer
RENEE DALY, Senior Program Assistant (until April 2025)

Reviewers

This *Reference Manual on Scientific Evidence* was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each publication as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the statement of task. The review comments and draft manuscripts remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this manual:
HELEN ADAMS, U.S. District Court for the Southern District of Iowa
BRUCE ALBERTS (NAS), University of California, San Francisco
PAOLA ARLOTTA (NAS, NAM), Harvard University
ORLEY ASHENFELTER (NAS), Princeton University
PAUL BARBADORO, U.S. District Court for the District of New Hampshire
JOHN BUTLER, U.S. Department of Commerce
CLEOPATRA CABUZ (NAE), Honeywell Industrial Safety (retired)
FRED CATE, Indiana University
VINTON CERF (NAS, NAE), Google
EDWARD CHENG, Vanderbilt University
ELLEN WRIGHT CLAYTON (NAM), Vanderbilt University
CARL CRANOR, University of California, Riverside
CHRIS FIELD (NAS), Stanford University
ELIZABETH GEDDES, Shihata & Geddes
CHARLES HAAS (NAE), Drexel University
SARA HUSTON, Ann & Robert H. Lurie Children's Hospital of Chicago
STEVEN HYMAN (NAM), Broad Institute of MIT and Harvard
RAYMOND JACKSON, U.S. District Court for the Eastern District of Virginia
KATHLEEN HALL JAMIESON (NAS), University of Pennsylvania
PAUL JANICKE, University of Houston
KAREN KAFADAR, University of Virginia
MARGARET BULL KOVERA, John Jay College of Criminal Justice
ERIC LANDER (NAS, NAM), Broad Institute of MIT and Harvard
JULIA LEIGHTON, Public Defender Service for the District of Columbia (retired)
PATRICK MALONE, Patrick Malone & Associates
RICHARD A. MESERVE (NAE), Carnegie Science
ROBERT MILLER, U.S. District Court for the Northern District of Indiana
DAVID OWEN, University of South Carolina (retired)

MARIA POLYAKOVA, Stanford University
ANTHONY PORCELLI, U.S. District Court for the Middle District of Florida
WILLIAM PRESS (NAS), The University of Texas at Austin
JED RAKOFF, U.S. District Court for the Southern District of New York
NATALIE RAM, University of Maryland
MARGARET RILEY, University of Virginia
VICTORIA ROBERTS, U.S. District Court for the Eastern District of Michigan (retired)
JULIE ROBINSON, U.S. District Court for the District of Kansas
BARBARA ROTHSTEIN, U.S. District Court for the Western District of Washington
JOELLEN RUSSELL, University of Arizona
JOHN SAMET (NAM), University of Colorado
NATHAN A. SCHACHTMAN, UB Greensfelder
FRANCIS SHEN, Harvard University
MICHAEL SIMON, U.S. District Court for the District of Oregon
WILLIAM SMITH, U.S. District Court for the District of Rhode Island
HAROLD SOX (NAM)**,** Dartmouth University (retired)
HOLDEN THORP, American Association for the Advancement of Science
WENDY WAGNER, The University of Texas at Austin
LYNN WINMILL, U.S. District Court for the District of Idaho
JOHN WIXTED, University of California, San Diego
DIANE WOOD, University of Chicago
DONALD WUEBBLES, University of Illinois
JAMES WYNN, U.S. Court of Appeals for the Fourth Circuit

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the *Reference Manual on Scientific Evidence*, nor did they see the final draft before its release. The review of this edition was overseen by Judge Jed S. Rakoff, U.S. District Court for the Southern District of New York, and Judge Kathleen M. O'Malley (retired), Kate O'Malley LLC. They were responsible for making certain that an independent examination of the manual was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests with the Committee on Science for Judges—Development of the Reference Manual on Scientific Evidence, Fourth Edition, the National Academies, and the Federal Judicial Center.

Foreword

I am pleased to introduce this fourth edition of the *Reference Manual on Scientific Evidence*. A joint product of the Federal Judicial Center and the National Academies of Sciences, Engineering, and Medicine, the manual has for three decades assisted judges in considering all manner of statistical and scientific evidence offered in litigation. In so doing, it has helped bring about better and fairer legal decisions.

It will surprise no one that science and law are today often intertwined. We live in an era of science and technology, with legal controversies reflecting that fact. And no sooner does one scientific field begin to become accessible to judges than another one emerges or fundamentally evolves. In the coming years, judges will confront lawsuits relating, for example, to artificial intelligence, climate science, and epidemiology. Even disputes not about science may involve statistical, survey, or other technical evidence of novel kinds.

Enter this manual. Judges typically are generalists, and they often lack extensive background in the sciences. They can learn, as they do on many subjects, through the adversary process, applying their critical faculties to the claims of parties. But sometimes it also helps to have a dispassionate guide. The manual is the product of close collaboration among highly respected scientists, engineers, judges, and lawyers. It delves into the scientific subjects that judges most often face. It explains scientific approaches and explores scientific uncertainties and limits. It aids in assessing the uses—and the misuses—of scientific and other technical evidence. The manual will hardly resolve every scientific issue arising in the law; that is not, and cannot be, its ambition. Yet case in and case out, the instruction that the manual offers in scientific principles and methods can improve the quality of judicial decision making.

In this way, the *Reference Manual on Scientific Evidence* exemplifies the benefits that can accrue from cooperation between those proficient in legal analysis and those expert in technical subjects. Judges, along with the lawyers who argue before them, will do their jobs better when they recognize what they can learn from the scientific community. And the law will become stronger as it further reflects sound science.

Elena Kagan
Associate Justice
United States Supreme Court

Summary Table of Contents

A detailed Table of Contents appears at the front of each reference guide.

Preface

This fourth edition of the *Reference Manual on Scientific Evidence* is the product of an ongoing collaboration between the Federal Judicial Center and the National Academies of Sciences, Engineering, and Medicine. The contributions of content by distinguished scholars from across disciplines have resulted in an impartial and reliable primary reference source on science for judges, which we set as our goal. We are grateful to the National Science Foundation and the Gordon and Betty Moore Foundation for the support that made this new edition possible.

Much has changed since the publication of the first edition of the *Reference Manual on Scientific Evidence* in 1994. Over the years, judges have become more comfortable in exercising a "gatekeeper" function over the admissibility of scientific evidence. Nevertheless, litigation is becoming ever more complex, and aggregate litigation now represents a larger share of the federal civil docket. Scientific, medical, and engineering evidence has also become more complex, and judges must often wrestle with unfamiliar and emerging science and technology. As a consequence, the rules governing admissibility have been revised to deal with emerging problems. For example, amendments to Rule 702 of the Federal Rules of Evidence were recently adopted to discourage testifying experts from overstating the value of evidence underlying such testimony. The dynamic and evolving presence of science in litigation necessitates engagement between the scientific and legal communities. The fourth edition of the *Reference Manual on Scientific Evidence* reflects this reality.

The primary (but not exclusive) audience for the manual is members of the judiciary who face challenging evidentiary issues related to science and technology. The manual is intended to assist judges in identifying issues commonly in dispute and to help judges reach an informed and reasoned assessment of those issues based on expert evidence that is faithful to the law and within the boundaries of scientifically sound knowledge. The manual is *not* intended to instruct judges concerning what evidence should be admissible or to establish minimum standards for acceptable scientific testimony.

In the new edition, all reference guides from the previous edition have undergone extensive revision or have been written by new authors. New guides have been added to address eyewitness identification, computer science, artificial intelligence, and climate science, along with a new foreword from Associate Justice of the U.S. Supreme Court Elena Kagan.

We are honored to have co-chaired the committee tasked with the development of this new edition of the *Reference Manual on Scientific Evidence*. Convened under the auspices of the National Academies' Committee on Science, Technology, and Law, the committee comprised six judges and seven experts from across science and engineering. All have given generously of their time, judgment, and wisdom to produce a volume that provides a crucial resource for the legal and scientific communities.

We are especially grateful to the authors who worked so diligently in drafting manuscripts for the committee's review and who responded so thoughtfully to the committee's and external reviewers' comments.

We are also grateful for the input of federal judges who responded to a December 2020 survey about how they used the manual and any changes and additions they thought would make it more helpful and to the co-chairs of the National Academies' 2021 workshop entitled *Emerging Areas of Science, Engineering, and Medicine for the Courts*.[1] The survey responses, the workshop, and 2019 and 2020 discussions[2] on science and the courts convened by Dr. Marcia McNutt, president of the National Academy of Sciences, identified many topics and issues that appear in this new edition of the manual.

Additionally, the committee received substantial encouragement from John Cooke, former director of the Federal Judicial Center, and Alan Thompkins, former deputy division director, acting division director, and acting deputy assistant director in the Social, Behavioral and Economic Sciences Directorate. The committee benefitted greatly from the work of Federal Judicial Center staff, particularly Meghan Dunn and Jason Cantone, and from the staff of the National Academies, particularly Steven Kendall, Renee Daly, and Dominic LoBuglio, and from editors Sally-Anne Cleveland, Nathan Dotson, Susanna McCrea, Clare O'Shea, and particularly Geoffrey Erwin.

Special commendation goes to Joe Cecil, project consultant, and Anne-Marie Mazza, project director and senior director of the Committee on Science, Technology, and Law of the National Academies, along with Beth Wiggins, Research Division director for the Federal Judicial Center. Their many contributions in shepherding the manual to completion have been invaluable.

Nancy D. Freudenthal
and Fred H. Gage
Committee Co-Chairs

1. *See* https://www.nationalacademies.org/science-for-courts-workshop.

2. Science in the Courts: Amicus Briefs and the Law, Feb. 2019; and Responsiveness of National Academies' Reports to the Needs of the Courts, Nov. 2020. These discussions were supported by the Ralph J. Cicerone Endowment Fund and the Annenberg Foundation Trust at Sunnylands.

The Admissibility of Expert Testimony

LIESA L. RICHTER AND DANIEL J. CAPRA

Liesa L. Richter, J.D., is George Lynn Cross Research Professor, Floyd & Martha Norris Chair in Law, University of Oklahoma College of Law, and Academic Consultant to the Judicial Conference Advisory Committee on Evidence Rules.

Daniel J. Capra, J.D., is Philip Reed Professor of Law, Fordham Law School, and Reporter to the Judicial Conference Advisory Committee on Evidence Rules.

CONTENTS

Introduction

The trial process operates to resolve disputed facts and claims through the presentation of evidence. The vast majority of evidence presented at trial comes in through the testimony of witnesses. Many of these witnesses are lay witnesses with personal knowledge of the facts and underlying events relevant to the issues in dispute. Often, however, the finder of fact requires assistance from witnesses who possess scientific, technical, or other specialized knowledge. Such witnesses offer expert opinion testimony that, when properly validated, helps jurors and judges resolve complex issues that are outside common knowledge and experience. Federal Rule of Evidence 702 regulates the admissibility of expert opinion testimony. Pursuant to Rule 702, the federal trial judge serves as a gatekeeper for expert opinion testimony, ensuring that only reliable expert testimony is considered by the factfinder.

The enactment, in 1975, of Federal Rule of Evidence 702, the interpretation of that rule by the U.S. Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[1] and the further amendments to Rule 702 as recently as 2023, have effected a gradual but significant change in how federal courts treat the admissibility of scientific evidence. Previously, a district court had to decide whether the proffered evidence was deduced from scientific principles generally accepted in the corresponding scientific community. Now, the district court must itself determine whether it is more likely than not that these principles are reliable and have been reliably applied to the facts of the case. This requires the court to become conversant with the science at issue—helped perhaps by this publication. At the same time, courts are cautioned not to prefer one valid scientific theory over another and to admit even differing expert opinions that satisfy the Rule 702 reliability standard. This reference guide examines how courts have tried to grapple with this difficult gatekeeper role.

Federal Rule of Evidence 702: An Overview

When a party objects to the admissibility of expert opinion testimony, Rule 702 requires a trial judge to analyze the basis of the challenge and to make certain findings before admitting the testimony.[2] Rule 104(a) of the Federal Rules of Evidence requires trial judges to decide preliminary questions regarding the admissibility of evidence by a preponderance of the evidence. Rule 104(a) applies to

1. 509 U.S. 595 (1993).

2. As with other evidence rules, a trial judge need not make findings under Rule 702 in the absence of an objection to expert opinion testimony. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

the trial judge's findings under Rule 702, as it does to the findings under most of the evidence rules.[3] To admit expert opinion testimony over objection, a trial judge must consider the following aspects of the testimony, if challenged. First, the judge must find the witness *qualified* to opine on the subject matter at issue. The judge must also find that the witness's scientific, technical, or other specialized knowledge *will help* the trier of fact. Before admitting the expert's opinion, the court must find that the opinion is based on *sufficient facts or data*. In addition, the court must determine that the witness's opinion is the product of *reliable principles and methods* and that the opinion reflects a *reliable application* of those principles and methods to the facts of the case. And a 2023 amendment to Rule 702 clarifies that the trial judge should determine that the expert witness does not overstate the conclusions that may be drawn from the methodology. The proponent of the testimony has the burden of satisfying the challenged aspects of Rule 702 by a preponderance of the evidence. Although Rule 702 characterizes opinion testimony based on scientific, technical, or other specialized knowledge as "expert" opinion testimony, some federal opinions caution trial courts to refrain from referring to witnesses as "experts" to avoid giving them undue influence with the factfinder.[4]

History of Rule 702

The Pre-Rules *Frye* Standard and Original Rule 702

In order to understand the current Rule 702 standard, it is helpful to be familiar with the history of expert opinion testimony in American trials and the standard

3. The Supreme Court, in *Bourjaily v. United States*, 483 U.S. 171 (1987), held that Rule 104(a) requires preliminary questions regarding the admissibility of evidence to be decided by the trial judge using a preponderance of the evidence standard.

4. *See, e.g.*, United States v. Maya, 966 F.3d 493, 505 (6th Cir. 2020) ("We have cautioned district courts not to declare a witness an expert in front of the jury."); United States v. Bartley, 855 F.2d 547, 552 (8th Cir. 1988) (noting that "[s]uch an offer and finding by the Court might influence the jury in its evaluation of the expert and the better procedure is to avoid an acknowledgment of the witnesses' expertise by the Court"). *See also* American Bar Association Civil Trial Practice Standard 14 (2007) ("The court should not, in the presence of the jury, declare that a witness is qualified as an expert or to render an expert opinion, and counsel should not ask the court to do so" because "use of the term 'expert' may appear to a jury to be a kind of judicial imprimatur that favors the witness."); Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("[T]here is much to be said for a practice that prohibits the use of the term 'expert' by both the parties and the court at trial. Such a practice 'ensures that trial courts do not inadvertently put their stamp of authority' on a witness' opinion, and protects against the jury's being 'overwhelmed by the so-called "experts."'") (*quoting* Hon. Charles R. Richey, Proposals to Eliminate the Prejudicial Effect of the Use of the Word "Expert" Under the Federal Rules of Evidence in Criminal and Civil Jury Trials, 154 F.R.D. 537, 559 (1994)).

of admissibility that applied before the Federal Rules. In 1923, in the case of *Frye v. United States*, the D.C. Circuit Court of Appeals held that polygraph evidence was properly excluded because it had not gained "general acceptance" in the scientific community.[5] Thereafter, the *Frye* "general acceptance" standard was widely adopted by federal and state courts alike to regulate the admissibility of expert testimony.[6] When the Federal Rules of Evidence took effect in 1975, the *Frye* "general acceptance" standard reigned. As originally enacted, Rule 702 provided simply that a qualified witness could testify if scientific, technical, or other specialized knowledge would "assist" the trier of fact to understand the evidence or to determine a fact in issue. It did not reference the *Frye* standard or give any clue to its continued application under the rules.

The Supreme Court's *Daubert* Trilogy

Until 1993, there was considerable controversy over whether Federal Rule of Evidence 702 incorporated the *Frye* "general acceptance" test. Then, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[7] the Supreme Court unanimously rejected the *Frye* "general acceptance" standard as the exclusive test for assessing the admissibility of scientific expert testimony under Rule 702.

Daubert was a case in which the plaintiffs claimed that the use of anti-nausea drug Bendectin during pregnancy caused birth defects in their babies. The district court granted summary judgment for the defendant after excluding the plaintiffs' expert on causation under *Frye*'s general acceptance standard. Because the expert's reliance on animal studies, chemical structure analysis, and reanalysis of epidemiological studies had not been generally accepted as methods for determining causation, the trial court excluded the expert's opinion testimony. On appeal, the Supreme Court scrutinized the language of Rule 702 and found nothing pointing to the continued application of the "general acceptance" standard. The Court further opined that a rigid "general acceptance" standard for the admissibility of scientific evidence would be at odds with the "liberal thrust" of the Federal Rules of Evidence and their otherwise flexible approach to opinion testimony. Therefore, the Court held that Rule 702 did not adopt the *Frye* "general acceptance" standard as the sole measure of the admissibility of expert opinion testimony.

5. 293 F. 1013 (D.C. Cir. 1923).

6. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 585 (1993) ("In the 70 years since its formulation in the *Frye* case, the 'general acceptance' test has been the dominant standard for determining the admissibility of novel scientific evidence at trial."). For a discussion of the admissibility of scientific evidence under the *Frye* standard, see Edward Imwinkelried et al., *Courtroom Criminal Evidence*, Vol. 1 §§ 608–613 (7th ed. 2022).

7. 509 U.S. 579 (1993).

Although it rejected "general acceptance" as the sine qua non of admissibility, the Court found limits on the admissibility of scientific expert opinion testimony implied by the terms "scientific" and "knowledge" in Rule 702. According to *Daubert*, a trial judge must ensure that scientific evidence is not only relevant, but *reliable*. Rule 702 requires the trial judge to act as a "gatekeeper" for scientific expert testimony, to ensure that it truly proceeds from "scientific . . . knowledge." The Court stated that an inference or assertion must be derived from the scientific method in order to qualify as scientific knowledge.

Essentially, the Court held that a trial judge must evaluate proffered scientific expert testimony to ensure that it is more likely than not reliable; concerns about the validity and reliability of expert testimony cannot simply be referred to the jury as questions of weight. The Court made this clear by noting that Rule 104(a) provides the standard for admitting challenged expert testimony. Rule 104(a) had previously been interpreted by the Court as requiring admissibility requirements to be established by a preponderance of the evidence.[8] Under *Daubert*, therefore, the judge must find it more likely than not that the expert's methods are reliable and that they are reliably applied to the facts at hand.

The move from *Frye* to *Daubert* dramatically changed the role of federal district judges. Under *Frye,* the trial judge could essentially count heads in the relevant scientific community to determine whether a particular methodology was generally accepted. Under *Daubert*, however, the decision about reliability cannot be completely outsourced to the scientific community; it is the trial judge who must examine the proffered scientific opinion testimony to determine whether it is reliable. Thus, a district judge faces the challenging task of understanding the science behind a proffered expert opinion sufficiently to assess its reliability. Chief Justice Rehnquist, dissenting from the majority's gatekeeper holding in *Daubert*, lamented that the majority had imposed on trial judges "either the obligation or the authority to become amateur scientists in order to perform that [gatekeeper] role."[9]

To assist judges in this new gatekeeping task, the majority in *Daubert* set forth a five-factor, nondispositive, nonexclusive, "flexible" test to be employed by the trial court under Rule 702 in determining the "validity" of scientific evidence. These factors—the famous "*Daubert* factors"—are:

1. whether the technique or theory can be or has been tested;
2. whether the theory or technique has been subject to peer review and publication;

8. The Court held that questions of admissibility under Rule 104(a) must be determined by a preponderance of the evidence in *Bourjaily v. United States*, 483 U.S. 171 (1987).

9. 509 U.S. at 601.

3. the known or potential rate of error;
4. the existence and maintenance of standards and controls; and
5. the degree to which the theory or technique has been generally accepted in the scientific community.

The factors described by the Court demand an objective assessment of an expert's technique or theory. With respect to testing, the Court explained that whether a scientific technique or theory has been or can be tested is ordinarily a key question in assessing its reliability because scientific methodology "is based on generating hypotheses and testing them to see if they can be falsified." The Court explained that publication is an important consideration because submission to the scrutiny of the scientific community increases the likelihood that substantive flaws in the methodology will be detected. Still, the Court recognized that publication is not an absolute requirement of admissibility given that well-grounded but innovative theories may not have been published. Finally, the Court explained that the general acceptance of a technique remains relevant to, although not dispositive of, admissibility. The Court explained that widespread acceptance can be an important factor in finding a methodology reliable.

The Ninth Circuit applied the gatekeeping standard on remand in the *Daubert* case, offering additional insights into the newly articulated test.[10] The court held that summary judgment was properly granted against the plaintiffs because their experts' testimony that Bendectin caused limb reduction was inadmissible under Rule 702. In so holding, the Ninth Circuit focused on "whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying." According to the court, "a scientist's normal workplace is the lab or the field, not the courtroom or the lawyer's office."[11] Thus, when expert opinion testimony is based on research conducted independent of litigation, this "provides important, objective proof that the research comports with the dictates of good science." The court reasoned that "independent research carries its own indicia of reliability, as it is conducted, so to speak, in the usual course of business and must normally satisfy a variety of standards to attract funding and institutional support."[12]

10. Daubert v. Merrell Dow Pharms., Inc., 43 F.3d 1311, 1317–19 (9th Cir. 1995).

11. *Id.*

12. *Id.* In casting doubt on research conducted in anticipation of litigation, the *Daubert on remand* court referred to research on issues of *general*, as opposed to specific, causation. In a toxic tort case, the plaintiff must show both that the toxic substance causes the harm suffered by the plaintiff (the issue of general causation) and that the toxic substance actually did cause the plaintiff's harm in the instant case (the issue of specific causation). *Daubert on remand* expressed concern about research into general causation that is conducted in anticipation of litigation. But an expert cannot be expected to conduct research on specific causation—the cause of a particular plaintiff's

The Ninth Circuit did not hold that an opinion that is *not* based on independent research is inadmissible. Instead, the court explained that "the party proffering it must come forward with other objective, verifiable evidence that the testimony is based on scientifically valid principles" if the expert's testimony is not based on independent research.[13] According to the Ninth Circuit, this means that "the experts must explain precisely how they went about reaching their conclusions and point to some objective source—a learned treatise, the policy statement of a professional association, a published article in a reputable scientific journal or the like—to show that they have followed the scientific method, as it is practiced by (at least) a recognized minority of scientists in their field."[14]

Applying these principles to the testimony of the plaintiffs' experts, the *Daubert on remand* court found that the *Daubert* standards had not been met. The experts' research on the question of general causation was prepared solely for purposes of Bendectin litigation. It had not been peer reviewed, nor had it even been deemed worthy of comment by the scientific community. Finally, the experts had not explained their methodology or verified it by reference to objective sources. The court concluded that it had been presented "with only the experts' qualifications, their conclusions and their assurances of reliability. Under *Daubert*, that's not enough."[15]

Federal opinions since *Daubert* have continued to articulate factors that may be important in assessing the admissibility of expert testimony, which are well summarized by the court in *In re Marriott Int'l, Inc., Customer Data Sec. Breach Litig.*: whether the expert will be testifying about matters that grow "naturally and directly out of research they have conducted independent of litigation, or whether they have developed their opinions expressly for purposes of testifying"; whether "the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion"; whether "the expert has adequately accounted for obvious alternative explanations"; whether "the expert is being as careful as he would be in his regular professional work outside his paid litigation consulting";

injury—outside the realm of litigation. It is only when the plaintiff discovers the injury that research into causation would begin, and it is also at that very point that litigation is anticipated. Indeed, every treating doctor who testifies to the plaintiff's injuries can be said to conduct her research in anticipation of litigation.

13. *See also* Wendell v. GlaxoSmithKline LLC, 858 F.3d 1227, 1235 (9th Cir. 2017) ("[T]he district court was wrong to put so much weight on the fact that the experts' opinions were not developed independently of litigation and had not been published. While independent research into the topic at issue is helpful to establish reliability, its absence does not mean the experts' methods were unreliable . . . [E]xpert testimony may still be reliable and admissible without peer review and publication. That is especially true when dealing with rare diseases that do not impel published studies.") (citations omitted).

14. *See Daubert*, 43 F.3d 1311, 1318 at n.9.

15. *Id.* at 1319.

and whether "the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give."[16]

The *Daubert* opinion sent some mixed messages and left some unanswered questions, however. It sent mixed messages by criticizing the rigidity of the *Frye* general acceptance standard as being too "austere" and by encouraging a "flexible" inquiry, while at the same time directing trial judges to ensure that an expert's opinion was properly derived from the "scientific method." It clearly announced the trial judge's obligation to find reliability by a preponderance of the evidence under Rule 104(a), but then noted that "[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence."[17] Of course, the trial methods referred to by the Court are the proper methods for attacking testimony—*once it has been found admissible*. But this language caused some to assume, mistakenly, that "shaky" expert opinion testimony should be admitted, with issues of reliability to be resolved by the jury as ones of weight.

Daubert also left open questions. The opinion very clearly described the standard for admitting "scientific" expert testimony, leaving open the questions of which testimony falls within this category, as well as the standard applicable to nonscientific expert opinion testimony. Finally, because the Supreme Court remanded the case to the lower courts, it did not apply the reliability standard announced in *Daubert* to a concrete set of facts, leaving the lower courts to apply the new standard in the first instance.

Some of these open questions were addressed by the Supreme Court soon after. The Court offered insight into the proper application of the *Daubert* reliability standard in *General Electric Co. v. Joiner*.[18] In *Joiner*, the plaintiff proffered experts to testify that his exposure to polychlorinated biphenyls (PCBs) promoted his development of small cell lung cancer. The district court excluded the plaintiff's experts as unreliable under *Daubert* and granted summary judgment for the defendants. On appeal, the Eleventh Circuit reversed, stating that "[b]ecause the Federal Rules of Evidence governing expert testimony display a preference for admissibility, we apply a particularly stringent standard of review to the trial judge's exclusion of expert testimony."[19] Under this stringent standard of review, the Eleventh Circuit reversed the exclusion of the plaintiff's experts.

On appeal, the Supreme Court first held that the abuse of discretion standard of review applies to a district court's rulings admitting or excluding expert testimony under Rule 702 and that the Eleventh Circuit had erred in applying a

16. *In re* Marriott Int'l, Inc., Customer Data Sec. Breach Litig., 602 F. Supp. 3d 767, 774 (D. Md. 2022).

17. 509 U.S. at 596.

18. 522 U.S. 136, 146 (1997).

19. Joiner v. Gen. Elec. Co., 78 F.3d 524 (11th Cir. 1996).

more rigid standard.[20] Thus, appellate review of a trial court's *Daubert* ruling proceeds in two steps. The appellate court reviews de novo whether a trial court correctly applied the *Daubert*/Rule 702 framework and reviews for abuse of discretion the trial court's ruling admitting or excluding expert testimony.[21] The Supreme Court in *Joiner* found no abuse of discretion in the trial court's exclusion of the plaintiff's experts as unreliable, emphasizing that a trial judge must ensure that an expert's *application* of her principles and methods to the facts of the case is also reliable. The experts relied on several studies, but the Court found that these studies did not support the experts' conclusions as to causation. In one study, infant mice were exposed to PCBs and contracted a different kind of cancer than that suffered by the plaintiff. Moreover, the studies could not be replicated in adult mice or in any other species. The experts' reliance on four epidemiological studies was likewise flawed. Two of the studies found no statistically significant connection between PCBs and cancer; one study did not mention PCBs; and the fourth study, which found a statistically significant connection, involved subjects who were exposed to a variety of other carcinogens. The Court concluded that there was an "analytical gap" between the experts' conclusion on causation and the studies on which they relied for their conclusion. The experts made no attempt to bridge this analytical gap by relying on any other objective or verifiable standards. The *Joiner* Court concluded that the expert testimony was merely "ipse dixit"—it is so because I am an expert and I say so. According to the Court, such testimony fails to fulfill the *Daubert* requirement that an expert's conclusions be based on an objective and verifiable methodology.

Justice Stevens dissented from this portion of the opinion, arguing that the majority had examined each study relied upon by the plaintiff's experts in a piecemeal fashion and concluded that the experts' opinions on causation were unreliable because no *one* study supported causation. Justice Stevens argued that a "weight of the evidence" approach to scientific reasoning, in which the experts relied on *all of the studies* taken together, as well as interviews of the plaintiff and an examination of his medical records to conclude that his exposure to PCBs promoted his development of lung cancer, was appropriate and acceptable under *Daubert*.[22]

Daubert stated that trial courts should focus "on principles and methodology, not on the conclusions that they generate." But the *Joiner* Court clarified that a trial court should not let an opinion pass through the gate of admissibility when there is a clear disconnect between the stated methods and the conclusion reached by the expert. The *Joiner* Court explained that "conclusions and

20. 522 U.S. at 143.

21. Kirk v. Clark Equip. Co., 991 F.3d 865, 872 (7th Cir. 2021).

22. *Id.* For a discussion of the "weight of the evidence" methodology, see section titled "Weight of the Evidence Analysis: Closing the Analytical Gap" below.

methodology are not entirely distinct from one another" because if an expert uses a reliable method and comes to an outlier conclusion, the gatekeeper can fairly assume that there is some disconnect in the expert's application of that methodology. According to the Court, an expert must not only be using a reliable methodology (like reviewing studies, as in *Joiner*); she must also be reliably applying that methodology to the facts at issue. As one court has stated, trial judges "may evaluate the data offered to support an expert's bottom-line opinions to determine if that data provides adequate support to mark the expert's testimony as reliable."[23]

Joiner did not authorize trial courts "to determine which of several competing scientific theories has the best provenance."[24] Further, "*Daubert* does not require that a party who proffers expert testimony carry the burden of proving to the judge that the expert's assessment of the situation is actually correct."[25] Qualified scientific experts applying reliable methodologies might still have reasonable differences of opinion. The proponent of the evidence must show only that the expert's conclusion has been arrived at in a scientifically sound and methodologically reliable fashion.

Another question left open by *Daubert* was whether its standards apply to expert testimony that does not purport to be scientifically based. In *Kumho Tire Co. v. Carmichael*,[26] the Court held that the *Daubert* reliability standard applies to nonscientific expert opinion testimony. *Kumho* involved a tire failure expert who would have testified that a tire exploded because it was defective. His mode of analysis was to investigate whether the failure might have been caused for any reason other than defect. He employed his own four-factor test to rule out other causes such as underinflation. He found that some of his self-selected factors indicated improper inflation of the tire. Still, the expert proposed to testify that improper inflation was not the cause of the tire failure. The trial judge excluded the expert's testimony for, among other reasons, failure to satisfy the *Daubert* standards. The judge found that the expert's methodology was subjective: it had not been peer reviewed; there was no indication of the rate of error; and there was no general acceptance of the expert's four-factor test for determining alternative causation.

On appeal, the Supreme Court reasoned that Rule 702 requires experts to testify on the basis of "knowledge," and that in *Daubert*, "the Court specified that it is the rule's word 'knowledge,' not the modifying words like 'scientific', that 'establishes a standard of evidentiary reliability.'" The Court also noted that "it would prove difficult, if not impossible, for judges to administer evidentiary

23. Ruiz-Troche v. Pepsi Cola of P.R. Bottling Co., 161 F.3d 77, 81 (1st Cir. 1998).

24. *Id*.

25. *See* Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("When a trial court, applying this amendment, rules that an expert's testimony is reliable, this does not necessarily mean that contradictory testimony is unreliable.").

26. 526 U.S. 137, 147–48 (1999).

rules under which a gatekeeping obligation depended upon a distinction between 'scientific' knowledge and 'technical' or 'other specialized' knowledge. There is no clear line that divides the one from the others."[27] The Court therefore found that the Rule 702 reliability standard applies to all expert opinion testimony—whether supported by scientific, technical, or other specialized knowledge.

The Court then examined whether a trial judge could assess nonscientific expert testimony using the specific factors that *Daubert* had listed as relevant to assessing the reliability of scientific expert testimony. The Court held that any or all of these factors might be used to evaluate nonscientific expert testimony, but found that the trial judge must have considerable leeway in deciding *how* to determine whether particular expert testimony is reliable:

> *Daubert*'s list of specific factors neither necessarily nor exclusively applies to all experts or in every case. Rather, the law grants a district court the same broad latitude when it decides *how* to determine reliability as it enjoys in respect to its ultimate reliability determination.[28]

The Court emphasized that the factors that a trial judge uses to determine reliability must be dependent upon the particular area of expertise that is being evaluated. The Court explained as follows:

> [W]e can neither rule out, nor rule in, for all cases and for all time the applicability of the factors mentioned in *Daubert*, nor can we now do so for subsets of cases categorized by category of expert or by kind of evidence. Too much depends upon the particular circumstances of the particular case at issue.
>
> At the same time . . . some of *Daubert*'s questions can help to evaluate the reliability even of experience-based testimony. In certain cases, it will be appropriate for the trial judge to ask, for example, how often an engineering expert's experience-based methodology has produced erroneous results, or whether such a method is generally accepted in the relevant engineering community. Likewise, it will at times be useful to ask even of a witness whose expertise is based purely on experience, say, a perfume tester able to distinguish among 140 odors at a sniff, whether his preparation is of a kind that others in the field would recognize as acceptable.[29]

The *Kumho* Court provided a bottom-line standard for employing the gatekeeping requirement in evaluating nonscientific expert testimony:

> The objective of that requirement . . . is to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field.[30]

27. *Id.* at 148.
28. *Id.* at 142.
29. 526 U.S. at 151.
30. *Id.* at 152.

Turning to the tire failure expert's proffered testimony in *Kumho*, the Court found no abuse of discretion in either the trial judge's use of specific *Daubert* factors or in the ultimate decision to exclude the testimony as insufficiently reliable. The expert's assumption—that alternative causes could be ruled out by looking at a number of specific factors identified solely by the expert—was essentially subjective and unsupported by any other tire failure expert. Moreover, the expert discounted evidence that under his own test indicated that the tire had been improperly inflated. The Court concluded as follows:

> Indeed, no one has argued that Carlson himself, were he still working for Michelin, would have concluded in a report to his employer that a similar tire was similarly defective on grounds identical to those upon which he rested his conclusion here. Of course, Carlson himself claimed that his method was accurate, but, as we pointed out in *Joiner*, "nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the ipse dixit of the expert."[31]

Thus, the expert was properly rejected because he could not show that he used the same intellectual rigor in reaching his conclusion as would be expected of him in his professional life outside the courtroom.

Amendments to Rule 702

Rule 702 has been amended since its original enactment to clarify the trial judge's gatekeeping role announced in *Daubert*. In 2000, the rule was amended to incorporate the teachings of the Supreme Court's *Daubert* trilogy. In 2011, Rule 702 was amended as part of the restyling project that was designed to make the rules "more easily understood and to make style and terminology consistent throughout the rules." In 2023, Rule 702 was amended again to emphasize the trial judge's obligation to find its requirements satisfied by a preponderance of the evidence and to highlight the problem of expert overstatement.[32] Because the amendments to Rule 702 have codified principles governing the admissibility of expert opinion testimony found in the caselaw, pre-amendment cases may remain instructive in applying Rule 702, as amended. Judges and litigants should exercise caution in relying on pre-amendment cases, however, because the 2023 amendment was designed to correct misapplications of the rule in some federal cases.[33]

31. *Id.* at 157.

32. The amendment took effect on December 1, 2023.

33. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment (". . . many courts have held that the critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a).").

The 2000 Amendment to Rule 702

In 2000, Rule 702 was amended for two purposes: 1) to codify, structure, and expand upon the Supreme Court trilogy of *Daubert*, *Joiner*, and *Kumho* and its extensive progeny; and 2) to provide an extensive Committee Note to assist trial courts in exercising the gatekeeping function allocated to them under *Daubert*.[34] The amendment retained the preexisting standards for qualifying an expert witness and for ensuring that expert opinion testimony "will help" the jury (which were part of the original rule and were placed in a new subdivision (a)). The amendment added subsections (b)–(d) to the structure of the rule, setting forth three reliability-based standards that must be met before a challenged expert's testimony can be admitted.

Rule 702: A Qualified Expert

Amended Rule 702 retained the requirement that an expert be qualified by "knowledge, skill, experience, training, or education." This flexible qualification standard allows courts to qualify expert witnesses on a variety of grounds and in many disciplines. For example, in *United States v. Protho*, a panel of the Seventh Circuit upheld the district court's decision to admit expert opinion testimony about the make, model, and year of a kidnapper's vehicle from surveillance videos despite the defendant's objection that the witness lacked the requisite expertise. The appellate court found that the witness's "position, engineering education, and nearly three decades at Ford make him abundantly qualified to opine on the appearance and identity of Ford's products."[35] In *Cedar Lodge Plantation, L.L.C. v. CSHV Fairway View I, L.L.C.*, the district court found the plaintiff's environmental expert unqualified to offer reliable expert testimony because his experience was related to the resolution of hazardous waste matters for commercial and industrial facilities, rather than sewage systems for apartment complexes or multi-family residential communities like the one at issue in the case.[36] But a panel of the Fifth Circuit, in an unpublished opinion, found that the expert had "extensive experience in analysis and evaluation of environmental

34. *See In re* Marriott Int'l, Inc., Customer Data Sec. Breach Litig., No. 19-MD-2879, 2022 WL 1323139, at *4 (D. Md. May 3, 2022) ("The advisory note to the 2000 amendments to Rule 702 ('2000 Advisory Note') is essential reading for judges and lawyers who undertake a *Daubert* analysis.").

35. United States v. Protho, 41 F.4th 812, 823 (7th Cir. 2022). *See also* United States v. Smith, 919 F.3d 825, 835 (4th Cir. 2019) ("twelve-year veteran of the FBI with substantial experience in drug and gang investigations (conducting controlled buys, interviewing drug dealers and gang members, using gang members as confidential sources, listening to thousands of recorded conversations, and so forth)" qualified as an expert in drug and gang terminology).

36. 753 F. App'x 191, 195–96 (5th Cir. 2018).

contaminants, the area in which he was offered as an expert, including experience working on sewage systems for residential neighborhood communities." The court held that the witness's lack of specialization in sewage facilities for multi-family residential units did not render him unqualified.

Rule 702(a): The Expert's Opinion "Will Help" the Trier of Fact

Rule 702 also requires that expert opinion testimony "will help" the trier of fact "to understand the evidence or to determine a fact in issue." "[T]he 'help' requirement [from Rule 702] is satisfied where the expert testimony advances the trier of fact's understanding to any degree."[37] The original Advisory Committee's note to Rule 702 quoted a law review article explaining the helpfulness requirement as follows:

> There is no more certain test for determining when experts may be used than the common sense inquiry whether the untrained layman would be qualified to determine intelligently and to the best possible degree the particular issue without enlightenment from those having a specialized understanding of the subject involved in the dispute.[38]

According to the original Advisory Committee's note, unhelpful opinions are "superfluous and a waste of time." In *Daubert*, the Supreme Court also characterized this requirement as one of relevance and "fit" between the opinion testimony and the dispute at issue.[39]

United States v. King offers an example of a court evaluating an expert's helpfulness. In *King,* the defendant in a pill mill case objected to the testimony of a medical expert regarding standards of practice in the field of pain management, guidelines for prescribing medications, and "red flags" that indicated potential painkiller abuse.[40] The Eighth Circuit affirmed the admission of the testimony even though the expert could not definitively state that any particular prescription was illegitimate. The court found that the expert's opinion on the general operation of the clinic "advanced the trier of fact's understanding of the clinical practices at [the clinics] and how they differed from ordinary medical facilities." With regard to experts relying on the scientific method, it is safe to assume that

37. *See, e.g.*, United States *ex rel.* Morsell v. Symantec Corp., 2020 U.S. Dist. LEXIS 54847, *12 (D.D.C. 2020) (quoting 29 Charles Alan Wright & Arthur R. Miller, Federal Practice & Procedure § 6264.1 (2015)).

38. *See* Fed. R. Evid. 702 advisory committee's note to 1973 amendment (quoting Ladd, *Expert Testimony*, 5 Vand. L. Rev. 414, 418 (1952)).

39. 509 U.S. 579, 591 (1993).

40. 898 F.3d 797, 805–06 (8th Cir. 2018).

if the testimony is relevant, it will be helpful to factfinders who are unschooled in science.

Rule 702(b): The Expert's Opinion Testimony Is Based on Sufficient Facts or Data

This subsection requires an expert to have a sufficient *foundation* for opinion testimony. It imposes a *quantitative* requirement, ensuring that an expert has enough basis supporting the opinion to which she testifies.[41] If an expert has engaged in insufficient research, or has ignored obvious factors, the opinion must be excluded under this prong of the test.[42] Put colloquially, an expert must do her homework before offering opinion testimony. A requirement that an expert have enough data to support an opinion flows naturally from the Supreme Court's demand for "reliable" opinion testimony in *Daubert*. While this foundation requirement was "well developed in the case law and in the experience of trial lawyers and judges" prior to the 2000 amendment, it was not expressly grounded in one of the federal rules.[43]

A good example of a faulty foundation is found in *Stephens v. Union Pacific Railroad*.[44] The plaintiff in that case suffered from mesothelioma. The plaintiff's father worked for Union Pacific when the plaintiff was a young child. The plaintiff claimed that he was exposed to asbestos fibers as a child from his father's work clothes. In an effort to defeat a summary judgment motion by the defendant, the plaintiff proffered two experts who concluded that the plaintiff's exposure to asbestos fibers from his father's work clothes was a substantial factor in causing his mesothelioma. A key premise underlying both experts' opinions was that the plaintiff was exposed to significant levels of asbestos as a result of his father's work. In affirming summary judgment for the defendant, the Ninth Circuit emphasized that neither expert had any information about the degree to which the plaintiff's father was exposed to asbestos fibers in his job, or about the plaintiff's own level of exposure, if any. Without that information, the court explained that the experts had insufficient basis to conclude that the plaintiff had

41. *See* Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("Subpart (1) of Rule 702 calls for a quantitative rather than qualitative analysis.").

42. *See, e.g., In re* Incretin-Based Therapies Prod. Liab. Litig., No. 21-55342, 2022 WL 898595, at *1 (9th Cir. Mar. 28, 2022) ("district court properly relied on the uncontested fact that Dr. Gale did not independently review studies that had been published between 2015 and Dr. Gale's final 2019 report, all of which found no causal relationship between liraglutide use and the development of pancreatic cancer").

43. Elcock v. Kmart Corp., 233 F.3d 734, 756, n.13 (3d Cir. 2000).

44. 935 F.3d 852 (9th Cir. 2019).

been exposed to asbestos with any regularity. Therefore, the experts' conclusions were based on insufficient facts or data.[45]

As explained by the Advisory Committee's note to the 2000 amendment, "[t]here has been some confusion over the relationship between Rules 702 and 703" and their respective roles in regulating the basis for an expert opinion. Rule 702(b) provides a *quantitative* requirement. It demands "sufficient" facts or data supporting an expert's opinion, without which the opinion would not be reliable. Rule 703 addresses the type or *quality* of the facts or data that may be relied on by an expert. It releases expert witnesses from the standard requirement of personal knowledge, allowing them to base opinions on "facts or data in the case that the expert has been made aware of or personally observed." It then addresses the situation in which an expert relies on facts or data that would be inadmissible in evidence under the Rules to form the basis of an opinion. Rule 703 provides that experts may rely on such inadmissible basis information so long as "experts in the particular field would reasonably rely on those kinds of facts or data in forming an opinion on the subject."[46] For example, in *Nicholson v. Biomet, Inc.*, a biomedical engineer offered an opinion based, in part, on medical literature and reports that constituted inadmissible hearsay.[47] Because biomedical engineers "unquestionably rely on such data from medical experts when designing medical devices compatible with the human body," the appellate court found no error in allowing the testimony.[48] Importantly, Rule 703 prohibits the proponent of the expert from disclosing inadmissible basis information to the fact-finder unless the "probative value in helping the jury evaluate the opinion substantially outweighs" its prejudicial effect. In sum, Rule 702 requires a sufficient quantity of facts or data supporting an expert opinion, while Rule 703 regulates the type of data upon which an expert may rely, as well as the disclosure of inadmissible basis to the jury by the proponent.

45. *Id.*; *see also* Taylor v. Bristol Myers Squibb Co., 93 F.4th 339, 348 (6th Cir. 2024) (expert who relies on one study and ignores contrary studies, for no justifiable reason, is relying on insufficient facts or data; court notes that the 2023 amendment to Rule 702 was intended to correct court decisions holding that the sufficiency of an expert's basis was a question of weight and not admissibility); Wasson v. Peabody Coal Co., 542 F.3d 1172 (7th Cir. 2008) (expert's testimony that lessee's coal price for calculating royalties was too low based on only a single customer; as such it was not based on sufficient facts or data and was properly excluded under Rule 702(b)).

46. *See, e.g.*, Nicholson v. Biomet, Inc., 46 F.4th 757 (8th Cir. 2022) ("Biomedical engineers, such as Truman herself, unquestionably rely on such data from medical experts when designing medical devices compatible with the human body. Thus, Truman correctly used medical reports and literature as contemplated by Rule 703 to support her opinion as a biomedical engineer on the design of the M2a Magnum.").

47. *Id.*

48. *Id.*

Rule 702(c): The Testimony Is the Product of Reliable Principles and Methods

This subsection was added to Rule 702 by the 2000 amendment to reflect the heart of the *Daubert* opinion in rule text. As the Court in *Daubert* held, the trial judge must act as a gatekeeper and must find that the principles and methods utilized by an expert to formulate an opinion are reliable.

United States v. Gissantaner offers an analysis of the reliable methodology requirement.[49] The defendant in that case was charged with the unlawful possession of a firearm by a convicted felon after a witness complained that the defendant was handling a gun and a gun was found in his roommate's drawer. The handle of the gun contained DNA that emanated from multiple sources and the prosecution sought to present testimony based on STRmix, a probabilistic genotyping software program. The testimony would have shown that the software program produced a 49 million to 1 likelihood ratio for the defendant's DNA, offering "very strong support" for the conclusion that the defendant was a contributor to the DNA sample on the handle of the gun. The district court excluded the STRmix testimony under *Daubert*, finding it insufficiently reliable as a methodology.

On appeal, the Sixth Circuit conducted a careful analysis of the reliability of STRmix as a method of identifying contributing sources to multisource DNA samples, and reversed. First, the court noted that the STRmix methodology was capable of being tested and had, in fact, been tested using lab-created multisource DNA samples to evaluate the ability of the software to include and exclude sources. Next, the court explained that STRmix has been subjected to peer review because it has been the subject of over 50 peer-reviewed scientific articles. The court noted that two of three studies performed on STRmix were conducted by individuals who were not connected to the development of the software. The court examined the error rate for the STRmix program, finding that STRmix accurately excluded noncontributors to a DNA source 99.1% of the time and produced extremely low likelihood ratios for the 1% of sources that it failed to exclude. The court also noted that a national association of forensic laboratories had developed standards regarding the use of probabilistic genotyping software that were used to minimize errors associated with STRmix and other similar software programs. Finally, the court examined the general acceptance of STRmix as a method for determining contributors to a multisource DNA sample. The court found that STRmix is the "market leader" in probabilistic genotype software that is utilized by the FBI and numerous government laboratories in the U.S. and abroad. After carefully examining each of

49. 990 F.3d 457, 468 (6th Cir. 2021).

the factors outlined in *Daubert*, the court found that STRmix was a reliable methodology under Rule 702(c).[50]

It will often be the case that experts in a particular field employ different methods. *Daubert* and Rule 702(c) do not authorize a trial judge to choose one competing methodology over another if both have been shown to be reliable. The question is whether the methodology employed by the expert is reliable; dispute in the field is a relevant consideration, but it is not dispositive. For example, in *Ruiz-Troche v. Pepsi Cola of P.R. Bottling Co.*,[51] a wrongful death action arising from a traffic accident, the defendant proffered an expert pharmacologist who concluded that the decedent had "snorted" at least 200 milligrams of cocaine thirty to sixty minutes prior to the accident. To reach this conclusion on dosage, the expert relied on a mathematical formula said to be supported by several published studies utilizing a "half-life technique." The trial court excluded this testimony on the ground that other pharmacologists used a different technique for determining dosage. But the appellate court found an abuse of discretion. The court recognized that the scientific literature relied on by the expert "does not make an open-and-shut case" because the studies indicate that the half-life of cocaine varies among individuals due to many factors. But the court stated that the trial judge "set the bar too high" in excluding the expert's testimony on the basis of these possible variables. The court noted that "*Daubert* neither requires nor empowers trial courts to determine which of several competing scientific theories has the best provenance. It demands only that the expert's conclusion has been arrived at in a scientifically sound and methodologically reliable fashion." In this case, the expert's opinion was "premised on an accepted technique" and "embodied a methodology that has significant support in the relevant universe of scientific literature."[52]

Rule 702(d): The Witness Has Applied the Principles and Methods Reliably to the Facts of the Case

This subsection was added to Rule 702 in 2000 to require a trial court "to scrutinize not only the principles and methods used by the expert, but also whether

50. *Id.*; *see also* United States v. Morgan, 45 F.4th 192 (D.C. Cir. 2022) (expert's cell tower location opinion supported by "drive testing" was based on reliable principles and methods where "drive testing technology has been relied upon, tested and reviewed for decades in the multibillion dollar wireless communications industry").

51. 161 F.3d 77, 85 (1st Cir. 1998).

52. *Id.*; *see also* Adams v. LabCorp of Am., 760 F.3d 1322 (11th Cir. 2014) (reversing district court's exclusion of plaintiffs' expert regarding the breach of a cytotechnologist's standard of care; expert testimony was admissible notwithstanding nonblinded review where expert used a well-established classification system to identify precancerous cells).

those principles and methods have been properly applied to the facts of the case."[53] This requirement was implicit in *Daubert*'s recognition of a gatekeeping role for the trial judge and explicit in *Joiner.* Although *Daubert* stated that trial courts should focus "on principles and methodology, not on the conclusions that they generate," the *Joiner* Court explained that "conclusions and methodology are not entirely distinct from one another." Thus, Rule 702 was amended in 2000 to make this reliable application requirement explicit in rule text. The Advisory Committee's note to the amendment explained:

> Under the amendment, as under *Daubert*, when an expert purports to apply principles and methods in accordance with professional standards, and yet reaches a conclusion that other experts in the field would not reach, the trial court may fairly suspect that the principles and methods have not been faithfully applied.

For example, in *McGill v. BP Exploration & Production, Inc.*, the plaintiff brought suit for illnesses he allegedly suffered as a result of his exposure to oil, dispersants, and decontaminants while participating in the cleanup of the *Deepwater Horizon* oil spill.[54] The district court excluded the plaintiff's causation expert, concluding that the expert lacked critical knowledge regarding the level of the alleged contaminants harmful to humans and the level of the plaintiff's exposure. The district court also excluded the opinion because the expert assumed that the plaintiff's illnesses were caused by exposure because of the temporal proximity between his injuries and the exposure.

A panel of the Fifth Circuit upheld the exclusion of the plaintiff's causation expert in an unpublished opinion. The expert relied on studies showing respiratory harm from exposure to the contaminants at issue. Still, the studies did not support the expert's conclusion that the *plaintiff's* illnesses resulted from his exposure because they did not show that the contaminants caused the illnesses from which the plaintiff suffered or the levels of exposure necessary to cause harm. The court explained:

> [T]here is a notable analytical gap between the facts he relies on and the conclusions he reaches. Dr. Stogner's deposition fails to address other potential causes of McGill's illness and the method by which he rules them out. Dr. Stogner fails to analyze the conditions of exposure McGill may have experienced. . . . Dr. Stogner was unable to answer questions regarding how much time McGill spent scooping up oil, how, where, or in what quantity Corexit was used, how exposure levels would change once substances were diluted in seawater, or how McGill's protective equipment would affect exposure.

53. Fed. R. Evid. 702 advisory committee's note to 2000 amendment.
54. 830 F. App'x 430, 433 (5th Cir. 2020).

Thus, the expert did not reliably apply his principles and methods to the facts of the plaintiff's case.[55]

Importantly, as explained in the Advisory Committee's note to the 2000 amendment, the trial court's obligation to determine that an expert has reliably applied her methodology does not authorize a judge to determine the *correctness* of an expert's conclusion. As the Ninth Circuit explained in *Elosu v. Middlefork Ranch Inc.*: "Although a district court may screen an expert opinion for reliability, and may reject testimony that is wholly speculative, it may not weigh the expert's conclusions or assume a factfinding role."[56] In that case, plaintiffs seeking recovery for a fire that burned their cabin to the ground proffered a report by a fire investigator. The expert opined that the fire was caused by an open-flame pilot light that ignited combustible vapors from an oil stain that had been applied to a wooden deck the day before the fire. Although the parties stipulated to the expert's qualifications and to the reliability of his methodology, the district court excluded his causation opinion for disagreement with its "ultimate conclusions." The trial court found that the substance of the opinion was "speculative, uncertain, and contradicted by multiple eyewitness accounts." The Ninth Circuit reversed the exclusion of the expert's testimony, finding that the trial court "disregarded much of the expert's scientific analysis, weighed the evidence on record, and demanded corroboration—factfinding steps that exceed the court's gatekeeping role." The court emphasized that "Rule 702 does not license a court to engage in freeform factfinding, to select between competing versions of the evidence, or to determine the veracity of the expert's conclusions at the admissibility stage."[57]

55. *Id.* In *McKiver v. Murphy-Brown, LLC*, the defendant objected to an expert's testimony that a DNA marker of hog feces could be found on the homes neighboring defendant's farm. 980 F.3d 937, 959 (4th Cir. 2020). The defendant contested the application of the DNA methodology in the case and argued that the expert's qualifications and reliance on the DNA marker could not support an opinion about odor at the neighboring property. The appellate court affirmed admission of the testimony, finding that "Pig2bac, upon which the report was based, has been used globally to demonstrate the traceability of swine fecal wastes and was applied in this case using protocols for sampling and analysis by a team of four Ph.D. holders, each with experience with field work, animal operations, and environmental health and engineering." The court further noted that the expert's DNA methodology was properly applied to support his opinion about tracing DNA compounds and that he did not offer an opinion on odor. *See also* United States v. Morgan, 45 F.4th 192 (D.C. Cir. 2022) (agent reliably applied "drive testing" methodology to approximate cell phone location when he confirmed that there were no changes in cell tower networks since the incident at issue before conducting the testing and subjected his draft analysis to peer review to detect and correct any application errors).

56. 26 F.4th 1017, 1020 (9th Cir. 2022). *See also* Kennedy v. Collagen Corp., 161 F.3d 1226, 1230 (9th Cir. 1998) (reversing exclusion of causation expert where studies upon which expert relied offered support for his conclusion that collagen injections caused lupus and stating "[j]udges in jury trials should not exclude [expert] testimony simply because they disagree with the conclusions of the expert").

57. *Elosu,* 26 F.4th at 1020.

The 2023 Amendment to Rule 702

Rule 702 was amended in 2023 to address two concerns.[58] First, *Daubert* clearly held that Rule 104(a) governs a trial court's admissibility decision under Rule 702. Rule 104(a) provides that the *court* must decide any preliminary question about whether evidence is admissible. In *Bourjaily v. United States*, the Supreme Court held that Rule 104(a) requires the proponent to demonstrate to the court that admissibility requirements are met by a preponderance of the evidence.[59] The Advisory Committee's note to the 2000 amendment to Rule 702 reiterated that this preponderance standard of proof applies to the trial judge's findings with respect to the admissibility of expert opinion testimony.[60] Despite this, many federal opinions have included incorrect characterizations of the applicable standard, stating that questions of the sufficiency of an expert's basis and the reliability of an expert's application are questions of weight for the jury. These cases are highly problematic because they largely abdicate the trial judge's gatekeeping role on the important questions of the sufficiency of an expert's basis and the reliability of an expert's application of a methodology to the facts of the case.[61]

Rule 702 was amended in 2023 to correct these misstatements and to clarify the applicable preponderance standard of proof.[62] The amendment expressly states that the proponent of expert opinion testimony must "demonstrate[] to the court" that the admissibility requirements of Rule 702 are "more likely than not" satisfied.[63]

The Advisory Committee's note to the 2023 amendment was careful to point out that some challenges to expert testimony will raise matters of weight rather than admissibility even under the proper Rule 104(a) standard:

> For example, if the court finds it more likely than not that an expert has a sufficient basis to support an opinion, the fact that the expert has not read every single study that exists will raise a question of weight and not admissibility. But

58. The amendment became effective December 1, 2023.

59. 483 U.S. 171 (1987).

60. *See* Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("the proponent has the burden of establishing that the pertinent admissibility requirements are met by a preponderance of the evidence").

61. Language incorrectly stating that the sufficiency of facts or data and reliable application are generally questions of weight and not admissibility can be found in *Loudermill v. Dow Chemical Co.*, 863 F.2d 566 (8th Cir. 1988); *Viterbo v. Dow Chemical Co.*, 826 F.2d 420 (5th Cir. 1987); and *Smith v. Ford Motor Co.*, 215 F.3d 713 (7th Cir. 2000).

62. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment ("the rule has been amended to clarify and emphasize that expert testimony may not be admitted unless the proponent demonstrates to the court that it is more likely than not that the proffered testimony meets the admissibility requirements set forth in the rule").

63. The Advisory Committee's note emphasizes that "there is no intent to raise any negative inference regarding the applicability of the Rule 104(a) standard of proof for other rules." Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

> this does not mean, as certain courts have held, that arguments about the sufficiency of an expert's basis always go to weight and not admissibility. Rather it means that once the court has found it more likely than not that the admissibility requirement has been met, any attack by the opponent will go only to the weight of the evidence.

Second, Rule 702 was amended in 2023 to address a common application problem in which an expert relies on reliable methods and principles to generate an opinion but overstates the conclusion that can reliably be drawn from the methodology. Rule 702(d) was amended to emphasize that each expert opinion must "stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology."[64] The Advisory Committee note to the amendment explains that, while the problem of expert overstatement can occur in any context, it is "especially pertinent to the testimony of forensic experts in both criminal and civil cases." For example, in *United States v. Machado-Erazo,* a cell phone location expert testified to the *precise* location of the defendant's phone.[65] The D.C. Circuit held that it was error to permit this testimony where current cell phone location technology can locate a phone only within a *general* area.[66] In contrast, the Seventh Circuit in *United States v. Protho* held that a forensic fiber analyst's testimony reliably applied forensic fiber analysis methods because the analyst "acknowledged that her results could not definitively identify fibers as coming from the same source" and "disclosed that fiber analysis can 'never' associate 'a single item to the exclusion of all others' and that consistency alone 'is not a means of positive identification.'"[67] The 2023 amendment to Rule

64. *Id.*

65. 901 F.3d 326 (D.C. Cir. 2018).

66. *Id.*; *see also* United States v. Hill, 818 F.3d 289, 295 (7th Cir. 2016) (declaring that cell site analysis expert testimony should include a "disclaimer" regarding accuracy; the expert should not "overpromise on the technique's precision or fail to account for its flaws"; cell site testimony was admissible where the agent testified that the defendant's phone records were "consistent" with him being at or near relevant locations at relevant times, but clarified that he could not state whether a phone was "absolutely at a specific address").

67. 41 F.4th 812, 821 (7th Cir. 2022). The reliability of forensic fiber comparison evidence has been subject to serious criticism. *See* Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward 162–63 (2009) (highlighting the lack of research determining the error rate of certain fiber-comparison methods). A concurring opinion in *Protho* found that the forensic fiber analysis should have been excluded because of the government's failure to show that it relied on reliable principles and methods under Rule 702(c). Still, the concurring opinion noted that the fiber expert attempted to refrain from overpromising on the capabilities of her analysis. *Protho*, 41 F.4th at 835 ("To Baloga's credit, she did not explicitly state that any two fibers 'matched' or claim to predict the likelihood that they came from a common source with any statistical precision. But she implied as much by repeatedly insisting during her testimony that one would not expect two random fibers to have all the common characteristics that she identified."). For a case involving expert overstatement outside the forensic context, see *United States v. Valencia-Lopez*, 971 F.3d 891 (9th Cir. 2020). The defendant in *Valencia-Lopez* contended that he was coerced to carry drugs for a cartel. The trial court allowed a law enforcement expert to testify that the

702(d) seeks to rein in expert overstatement by directing trial judges to determine that an expert's proffered conclusions "reflect[] a reliable application of the principles and methods to the facts of the case."

Procedures for Resolving Rule 702 Motions and Objections

Trial courts enjoy significant discretion in fashioning procedures for determining the admissibility of expert opinion testimony in a given case. As the Supreme Court explained in *Kumho Tire*:

> The trial court must have the same kind of latitude in deciding *how* to test an expert's reliability, and to decide whether or when special briefing or other proceedings are needed to investigate reliability, as it enjoys when it decides *whether or not* that expert's relevant testimony is reliable. Our opinion in *Joiner* makes clear that a court of appeals is to apply an abuse-of-discretion standard when it "review[s] a trial court's decision to admit or exclude expert testimony." 522 U.S. at 138–139, 118 S. Ct. 512. That standard applies as much to the trial court's decisions about how to determine reliability as to its ultimate conclusion. Otherwise, the trial judge would lack the discretionary authority needed both to avoid unnecessary "reliability" proceedings in ordinary cases where the reliability of an expert's methods is properly taken for granted, and to require appropriate proceedings in the less usual or more complex cases where cause for questioning the expert's reliability arises.[68]

Accordingly, trial judges may decide *Daubert* motions at trial based on voir dire of a proffered expert witness or before trial, based on briefing and supporting documentation or after lengthy *Daubert* hearings involving testimony from proffered expert witnesses. Whether to hold a *Daubert* hearing is within the discretion of the court.[69]

Although the trial court's gatekeeper review is often conducted through a *Daubert* hearing, a hearing is not required when the evidentiary record pertinent to the expert opinions is already well developed. For example, in *Miller v. Baker Implement Co.*,[70] the court found that the trial judge did not abuse discretion in excluding the plaintiff's expert testimony without a *Daubert* hearing. It noted

likelihood of a cartel using a coerced courier was "almost nil." The Ninth Circuit found that this opinion was erroneously admitted under Rule 702 because there was no showing of how the expert's experience could lead to a conclusion of such certainty.

68. 526 U.S. 137, 152 (1999) (emphasis in original).

69. *See* Fed. R. Evid. 702 advisory committee's note to 2000 amendment (noting that the rule "makes no attempt to set forth procedural requirements for exercising the trial court's gatekeeping function over expert testimony").

70. 439 F.3d 407, 412 (8th Cir. 2006).

that the plaintiff submitted affidavits, a detailed explanation of the proposed expert testimony, and a legal memorandum addressing the expert evidence issues. The court noted that while *Daubert* hearings "may be necessary in some cases, the basic requirement under the law is that parties have an opportunity to be heard before the district court makes its decisions."[71]

Daubert Motions and Summary Judgment

In civil cases, objections to expert testimony under Rule 702 are often made in the context of a motion for summary judgment. Rule 56 of the Federal Rules of Civil Procedure requires that material cited to support or dispute a fact in the summary judgment context can be "presented in a form that would be admissible in evidence." Therefore, a court must make admissibility determinations to resolve a motion for summary judgment. Defendants often file a *Daubert* motion in conjunction with a motion for summary judgment. They argue that plaintiff's proffered expert opinion testimony fails to satisfy Rule 702 and that, without a necessary expert opinion to support the plaintiff on a crucial issue, the defendant is entitled to summary judgment. As the court explained in *Cortes-Irizarry v. Corporation Insular*,[72] "[i]f proffered expert testimony fails to cross *Daubert*'s threshold for admissibility, a district court may exclude that evidence from consideration when passing upon a motion for summary judgment." The court in that case cautioned against trying to "gauge the reliability of expert proof on a truncated record," however. Although the trial court enjoys significant discretion in fashioning proper procedures for evaluating *Daubert* motions, courts should ensure that the proponent of the challenged expert testimony has an adequate opportunity to develop a record supporting admissibility at the summary judgment stage. A trial court may even wish to require a full presentation of the expert testimony, and rebuttal, to facilitate the admissibility determination. Courts have displayed considerable ingenuity in devising mechanisms that permit *Daubert* rulings to be made in conjunction with motions for summary judgment.[73]

71. *Id.*; *see also, e.g.,* McKiver v. Murphy-Brown, LLC, 980 F.3d 937 (4th Cir. 2020) (noting that the trial court "was entitled to rely on the parties' materials without requiring further submissions or a *Daubert* hearing" and that the trial court's ruling from the bench "reflected that it considered the parties' arguments and briefing").

72. 111 F.3d 184, 188 (1st Cir. 1997).

73. *See, e.g.*, Padillas v. Stork-Gamco, Inc., 186 F.3d 412, 418 (3d Cir. 1999) ("An *in limine* hearing will obviously not be required whenever a *Daubert* objection is raised to a proffer of expert evidence. Whether to hold one rests in the sound discretion of the district court. But when the ruling on admissibility turns on factual issues, as it does here, at least in the summary judgment context, failure to hold such a hearing may be an abuse of discretion."); *In re* Paoli R.R. Yard PCB Litig., 35 F.3d 717, 736, 739 (3d Cir. 1994) (discussing the use of in limine hearings); Claar v.

Credibility Determinations and Rule 702 Motions

As discussed above, trial judges must find the requirements of Rule 702 satisfied by a preponderance of the evidence before passing expert opinion testimony on to the jury. Questions sometimes arise about the judge's ability to make credibility determinations in performing this function. At first glance, it would seem impermissible for a trial judge to make a credibility assessment because it would encroach on the jury's fundamental role. But there are some special credibility determinations that may be necessary in evaluating Rule 702's reliability requirements.

The court considered the complex relationship between expert credibility and reliability in *Elcock v. Kmart Corp.*[74] The trial judge in *Elcock* held a *Daubert* hearing and determined that one of the plaintiff's experts did not pass the reliability threshold. The judge relied, in part, on the fact that the expert had engaged in criminal acts involving fraud, although this fraudulent activity was not in any way related to the expert's professional life. The Third Circuit found the trial court's reliance on the expert's prior bad acts and consideration of the expert's general credibility to be error. Still, the court made clear that a district judge is required to make some credibility choices as the Rule 702 gatekeeper:

> Consider a case in which an expert witness, during a *Daubert* hearing, claims to have looked at the key data that informed his proffered methodology, while the opponent offers testimony suggesting that the expert had not in fact conducted such an examination. Under such a scenario, a district court would necessarily have to address and resolve the credibility issue raised by the conflicting testimony in order to arrive at a conclusion regarding the reliability of the methodology at issue.[75]

While the court concluded that some credibility determinations would have to be made at a *Daubert* hearing, it emphasized that those determinations should be limited to testimony about how the expert reached her opinion, as opposed to witness credibility more generally. Thus, a distinction must be drawn between credibility evidence that bears directly on the expert's methods and application, and evidence on credibility that is more general.[76]

Burlington N. R.R., 29 F.3d 499, 502–05 (9th Cir. 1994) (discussing the district court's technique of ordering experts to submit serial affidavits explaining the reasoning and methodology underlying their conclusions).

74. 233 F.3d 734, 750–51 (3d Cir. 2000).

75. *Id.*

76. *See* Adams v. LabCorp of Am., 760 F.3d 1322 (11th Cir. 2014) (issues of potential bias as a result of nonblinded review and expert's philosophical bent were matters of weight for the jury); *see also* Edward J. Imwinkelreid, *Trial Judges—Gatekeepers or Usurpers? Can the Trial Judge Critically*

If the attack on credibility has nothing to do with the expert's methods, but only with the expert's general character for truthfulness or bias, the issue of credibility should be left to the jury—the opponent can bring impeachment evidence before the jury by way of cross-examination as with any witness. As applied to the facts of *Elcock*, the credibility evidence should not have been used by the trial court because it related to acts of dishonesty and fraud completely outside the expert's work in the particular case.[77]

Important Procedural Takeaways

Whether or not the trial court holds a *Daubert* hearing, it is important to keep in mind three procedural points regarding the gatekeeper function:

- The trial court must find that the proponent has or has not shown by a preponderance of the evidence that the expert's testimony satisfies the requirements of Rule 702. The trial court cannot simply refuse to decide whether the expert relied on sufficient facts and data to support her opinion, employed reliable methods and principles, and applied those methods reliably to the facts of the case at hand. Leaving these matters "to the jury" constitutes an abdication of the trial court's gatekeeper role under Rule 104(a). For example, the trial court in *McClain v. Metabolife Int'l, Inc.*[78] held a *Daubert* hearing regarding two witnesses who were providing opinions on pharmacology and chemistry. After the hearing, the trial court essentially washed its hands of the admissibility determination and sent both witnesses to the jury. The trial court explained that it did not "pretend to know enough to formulate a logical basis for a preclusionary order that would necessarily find, as a matter of law, that these witnesses cannot express to a jury the opinions they articulated to the court."

Assess the Admissibility of Expert Testimony Without Invading the Jury's Province to Evaluate the Credibility and Weight of the Testimony, 84 Marq. L. Rev. 1, 39 (2000) (stating that an expert's prior dishonesty or misconduct should not be considered by the court when the acts are wholly unrelated to the expert's use of a particular methodology, but that a court should take such dishonesty or misconduct into account when the nexus between the acts and the expert's methodology is more direct, e.g., when the prior dishonest acts involve fraud committed in connection with the earlier phases of a research project that serves as the foundation for the expert's proffered opinion).

77. *See also,* Cruz-Vazquez v. Mennonite Gen. Hosp. Inc., 613 F.3d 54 (1st Cir. 2010) (error to exclude expert because he was biased in favor of plaintiffs in medical cases and was generally affiliated with plaintiffs' lawyers; those considerations are for the jury in assessing the weight of the expert's testimony).

78. 401 F.3d 1233, 1239 (11th Cir. 2005).

The court of appeals found that the trial court had made an error of law by having "essentially abdicated its gatekeeping role. Although the trial court conducted a *Daubert* hearing, and both witnesses were subject to a thorough and extensive examination, the court ultimately disavowed its ability to handle the *Daubert* issues. This abdication was in itself an abuse of discretion."[79]

- The trial court must make findings on the record and set forth its reasoning on the record as well. Expressly stating that the court has or has not found the Rule 702 requirements satisfied by a preponderance of the evidence can demonstrate that the proper standard has been applied. Without such findings, there is no way for an appellate court to determine whether the gatekeeping function was properly exercised.[80]
- The gatekeeping function has been exercised flexibly by some courts in the context of bench trials. While Rule 702 certainly applies in bench trials, some courts have found that the gatekeeping function can be applied more flexibly because a trial judge serving as factfinder can be trusted to give dubious expert testimony the weight it deserves.[81]

Recurring Issues in the Application of Rule 702

There are several recurring issues bearing on the reliability of a scientific expert's testimony that courts have confronted in deciding motions under Rule 702. None of the issues discussed below is necessarily dispositive, but each has been considered in the cases.

79. *Id.*

80. *See, e.g.*, Certain Underwriters at Lloyd's, London v. Axon Pressure Prods. Inc., 951 F.3d 248 (5th Cir. 2020) (abuse of discretion to exclude expert testimony by a plenary order that offered no reasons for the ruling); United States v. Yeley-Davis, 632 F.3d 673 (10th Cir. 2011) (it was error to permit an agent to testify about how cell phone towers work without making findings on the record that the witness was qualified and his opinions were reliable); Smith v. Jenkins, 732 F.3d 51 (1st Cir. 2013) (remand required given the absence of any findings on the record regarding the reliability of the expert's opinion).

81. *See, e.g.*, Doe v. Hamilton Cnty. Bd. of Educ., 392 F.3d 840, 852 (6th Cir. 2004) (noting that the gatekeeper function was "designed to protect juries and is largely irrelevant in the context of a bench trial"); United States v. Brown, 415 F.3d 1257, 1268 (11th Cir. 2005) (Admissibility standards of Rule 702 "are more relaxed in a bench trial situation, where the judge is serving as factfinder and we are not concerned about dumping a barrage of questionable scientific evidence on a jury. . . . There is less need for the gatekeeper to keep the gate when the gatekeeper is keeping the gate only for himself."). *But see* UGI Sunbury v. Permanent Easement for 1.7575 Acres, 949 F.3d 825 (3d Cir. 2020) (finding error in admitting speculative expert testimony, and expressing skepticism about the cases stating that the gatekeeping function is relaxed in bench trials).

Extrapolation and the "Analytical Gap"

In forming their opinions, some experts start from an accepted scientific fact or premise but draw a conclusion that is not necessarily supported by the accepted premise. These experts have sometimes relied on logical analysis to go from premise to conclusion. But others have simply made an unsupported leap. If the process from premise to conclusion is not itself consistent with the scientific method or other well-accepted principles, then courts tend to exclude the testimony as based on unreliable and improper extrapolation.

For example, improper extrapolation occurs when an expert concludes, without supporting research, that a substance that causes one harm also causes a different harm. In *Lust v. Merrell Dow Pharmaceuticals, Inc.*,[82] the question was whether the plaintiff's birth defect, hemifacial microsomia, was caused by his mother's use of the drug Clomid. The expert based his conclusion to that effect on epidemiological studies that showed a link between Clomid and other types of birth defects. He concluded that because Clomid is capable of causing other birth defects, it also caused the plaintiff's hemifacial microsomia. The court explained that the expert "could have convinced the district court that his methodology was nevertheless scientific if he had explained how he went about reaching his conclusions and had pointed to an objective source demonstrating that his method and premises were generally accepted by or espoused by a recognized minority of teratologists." The court affirmed the trial court's exclusion of the expert's opinion, however, because he failed to do so, finding that the expert could not simply rely on epidemiological studies regarding the connection between Clomid and other birth defects to reach a conclusion about the plaintiff's distinct birth defect. The court found that the doctor's testimony "was influenced by litigation-driven financial incentive," and the doctor's premise—that a positive association between a drug and some birth defects indicates an association with other birth defects—was not recognized by even a minority of scientists.[83]

Red flags are also raised when a scientific expert purports to testify on the basis of a methodology that she transfers from one arena to a completely different area of inquiry—at least if there is no independent research supporting the

82. 89 F.3d 594, 597 (9th Cir. 1996).

83. *See also* Mitchell v. Gencorp., 165 F.3d 778, 782 (10th Cir. 1999) (experts could not testify that the plaintiff's exposure to chemicals similar to benzene would affect the body in the same way as benzene; without "scientific data supporting their conclusions that chemicals similar to benzene cause the same problems as benzene, the analytical gap in the experts' testimony is simply too wide for the opinions to establish causation"); *In re* Nexium Esomeprazole, 662 F. App'x 528, 530 (9th Cir. 2016) (opinion of orthopedic surgeon that drug Nexium could cause reduced bone mineral density and related fractures properly excluded where he "did not adequately explain how he inferred a causal relationship from epidemiological studies that did not come to such a conclusion themselves").

transfer. Thus, in *Braun v. Lorillard Inc.*,[84] the plaintiff sought to prove that the decedent's mesothelioma was caused by smoking cigarettes with a filter made with crocidolite asbestos. The decedent's lung tissue was tested for asbestos fibers, using the standard methodologies for the testing of human tissue of "bleach digestion" and "low temperature plasma ashing." No crocidolite fibers were found. The plaintiffs then retained an expert whose specialty was testing for asbestos in *building materials* to conduct tests on the decedent's lung tissue. This expert was unaware of the methodologies ordinarily employed in testing human tissue. He used the same test that he used on building materials to test the decedent's lung tissue, known as "high temperature ashing," and found crocidolite fibers in the tissue. The expert stated that high temperature ashing was as usable on tissue as on bricks—though he had never conducted such a test on tissue before this litigation. He also admitted that the high temperature could alter the chemistry of the sample, in which case it would be impossible to tell whether asbestos fibers were crocidolite or some other kind. But he asserted that his method was far more likely to produce a false negative than a false positive.

The Seventh Circuit, in an opinion by Judge Posner, held that the expert's testimony was properly excluded as unscientific under *Daubert*. It made the following points:

> A judge or jury is not equipped to evaluate scientific innovations. If, therefore, an expert proposes to depart from the generally accepted methodology of his field and embark upon a sea of scientific uncertainty, the court may appropriately insist that he ground his departure in demonstrable and scrupulous adherence to the scientist's creed of meticulous and objective inquiry. To forsake the accepted methods without even inquiring why they are the accepted methods—in this case, why specialists in testing human tissues for asbestos fibers have never used the familiar high temperature ashing method—and without even knowing what the accepted methods are, strikes us . . . as irresponsible.[85]

Judge Posner provided a bottom line: *Daubert* seeks to curtail "the hiring of reputable scientists, impressively credentialed, to testify for a fee to propositions that they have not arrived at through the methods that they use when they are doing their regular professional work rather than being paid to give an opinion helpful to one side in a lawsuit."[86]

All this is not to say that learning from one subject matter can never be used to form a conclusion as to a different subject matter, or that premises established in one context are completely irrelevant in another. Extrapolation by an expert is permissible when justified by scientific support or reasoning. An example of permissible extrapolation by a scientific expert arose in *Kennedy v. Collagen*

84. 84 F.3d 230, 234 (7th Cir. 1996).
85. *Id.* at 235.
86. *Id.*

Corp.[87] The plaintiff claimed that her use of the facial product Zyderm caused her to contract lupus, an autoimmune disorder. The plaintiff's expert rheumatologist testified to causation on the basis of peer-reviewed articles, clinical trials, product studies, an investigation conducted by the Texas Department of Health, a physical examination of the plaintiff, laboratory tests, and the temporal proximity between the plaintiff's use of Zyderm and the onset of her autoimmune disorder. None of the studies upon which the expert relied indicated that Zyderm caused lupus; but they did establish a connection between Zyderm and autoimmune disorders that were similar to lupus. Ordinarily, an analytical gap would arise if the expert were to conclude that Zyderm causes lupus simply because it causes disorders that are similar to lupus. However, the expert filled the analytical gap by relying on the finding, established in both peer-reviewed publications and clinical studies, that Zyderm induces the body to produce the *same autoimmune antibodies* that are the hallmark of lupus. The expert was not relying solely on causation of other autoimmune diseases, but also on the fact that the manner of causation was the same for lupus and for other autoimmune diseases that *were* linked to Zyderm. The court concluded that "Dr. Spindler set forth the steps he took in arriving at his conclusion in his deposition. Dr. Spindler's analogical reasoning was based on objective, verifiable evidence and scientific methodology of the kind traditionally used by rheumatologists. That is precisely what *Daubert* requires."[88]

Weight of the Evidence Analysis: Closing the Analytical Gap

In *Joiner*, the Supreme Court concluded that there was an "analytical gap" between the experts' conclusion on causation and the studies upon which they relied. As discussed above, Justice Stevens dissented, arguing that the majority had examined each study relied upon by the experts in a piecemeal fashion and had concluded that the experts' opinions on causation were unreliable because no *one* study supported causation. Justice Stevens argued that a "weight of the evidence" approach to scientific reasoning, in which the experts relied upon *all of the studies* taken together, as well as interviews of the plaintiff and an examination of his medical records to conclude that his exposure to PCBs promoted his development of lung cancer, was appropriate and acceptable under *Daubert*.[89] The majority did not address this argument directly, essentially finding insufficient support for the experts' conclusion in the studies and data relied upon in that case.

87. 161 F.3d 1226, 1230 (9th Cir. 1998).
88. *Id.*
89. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 152–54 (1997).

A "weight of the evidence" process of scientific reasoning has been found sufficient to support an expert's conclusion on general causation under Rule 702 since the *Joiner* decision. Weight of the evidence methodology allows a scientific expert to reason to a conclusion even though no single report or study exists that independently supports that conclusion. As the First Circuit described it in *Milward v. Acuity Specialty Products Group, Inc.*,[90] the weight of the evidence approach to making causal connections "is a mode of logical reasoning often described as 'inference to the best explanation' in which the conclusion is not guaranteed by the premises."[91]

The plaintiff in *Milward* brought negligence claims against chemical companies alleging that the plaintiff's rare type of leukemia, acute promyelocytic leukemia (APL), was caused by his routine workplace exposure to benzene-containing products. The district court excluded the plaintiff's general causation expert, finding that his proffered testimony that benzene exposure can cause APL lacked sufficient demonstrated scientific reliability. On appeal, a panel of the First Circuit reversed, finding that the plaintiff's expert opinion on general causation satisfied Rule 702. The court noted that the plaintiff's general causation expert relied on a "weight of the evidence" approach in concluding that benzene was capable of causing APL.

The court explained that this approach involves an inference to the best explanation that can be thought of as involving six general steps:

> The scientist must (1) identify an association between an exposure and a disease, (2) consider a range of plausible explanations for the association, (3) rank the rival explanations according to their plausibility, (4) seek additional evidence to separate the more plausible from the less plausible explanations, (5) consider all of the relevant available evidence, and (6) integrate the evidence using professional judgment to come to a conclusion about the best explanation.[92]

The court acknowledged that this method of reasoning depends upon the use of "scientific judgment" but explained: "[t]he fact that the role of judgment in the weight of the evidence approach is more readily apparent than it is in other methodologies does not mean that the approach is any less scientific."[93] The court noted that the weight of the evidence approach is commonly used by scientists, and opined that "[n]o serious argument can be made that the weight of the evidence approach is inherently unreliable" under Rule 702.

The court found that the question to be resolved by a trial court is whether the approach has been properly applied in a particular case. The *Milward* court found that the district court had erred in excluding the expert's opinion

90. 639 F.3d 11, 17–19 (1st Cir. 2011).
91. *Id.*
92. *Id.*
93. *Id.*

testimony on general causation where the expert had derived his opinion "from the accumulation of multiple scientifically acceptable inferences from different bodies of evidence" and clearly employed the "same level of intellectual rigor" that he did in his academic work. The *Milward* court found that the district court erred in treating "the separate evidentiary components of Dr. Smith's analysis atomistically":

> In Dr. Smith's weight of the evidence approach, no body of evidence was itself treated as justifying an inference of causation. Rather, each body of evidence was treated as grounds for the subsidiary conclusion that it would, if combined with other evidence, support a causal inference. The district court erred in reasoning that because no one line of evidence supported a reliable inference of causation, an inference of causation based on the totality of the evidence was unreliable. The hallmark of the weight of the evidence approach is reasoning to the best explanation for all of the available evidence.[94]

The First Circuit explained that the weight of the evidence approach may be particularly appropriate in cases like the plaintiff's because "the rarity of APL and difficulties of data collection in the United States make it very difficult to perform an epidemiological study of the causes of APL that would yield statistically significant results."[95] Thus, the court held that the trial court abused its discretion by concluding that the expert's inference of causation based on the totality of the evidence was unreliable "because no one line of evidence supported a reliable inference of causation."[96]

94. *Id.* (citations omitted).

95. *Id.*

96. *See also In re* Abilify (Aripiprazole) Prod. Liab. Litig., 299 F. Supp. 3d 1291, 1311 (N.D. Fla. 2018) ("This 'weight of the evidence' approach to analyzing causation can be considered reliable, provided the expert considers all available evidence carefully and explains how the relative weight of the various pieces of evidence led to his conclusion."); Waite v. AII Acquisition Corp., 194 F. Supp. 3d 1298, 1313 (S.D. Fla. 2016) ("This weight-of-the-evidence approach to causation requires consideration of all available scientific evidence, including epidemiology, toxicology (animal studies), cellular studies (in vitro), and molecular biology and has been validated by the courts."). *But see In re* Zoloft (Sertraline Hydrochloride) Prod. Liab. Litig., 858 F.3d 787, 796–97 (3d Cir. 2017) ("Here, we accept that the Bradford Hill and weight of the evidence analyses are generally reliable. We also assume that the 'techniques' used to implement the analysis (here, meta-analysis, trend analysis, and reanalysis) are themselves reliable. However, we find that Dr. Jewell did not 1) reliably apply the 'techniques' to the body of evidence or 2) adequately explain how this analysis supports specified Bradford Hill criteria. Because '*any* step that renders the analysis unreliable under the *Daubert factors renders the expert's testimony inadmissible*,' this is sufficient to show that the District Court did not abuse its discretion in excluding Dr. Jewell's testimony.") (citations omitted); *In re* Nexium Esomeprazole, 662 F. App'x 528, 530 (9th Cir. 2016) ("At best, Dr. Bal analyzed three of the nine Bradford Hill factors that guide scientists in drawing causal conclusions from epidemiological studies."). For a detailed analysis of the *Milward* decision and the weight of the evidence approach to scientific reasoning, see *Symposium: Toxic Tort Litigation: After Milward v. Acuity Products*, 3 Wake Forest J.L. & Pol'y 1 (2013).

Reliance on Anecdotal Evidence

Courts have generally found that an expert who bases a scientific opinion solely on one or even a few case studies has an insufficient basis for reliable testimony under Rule 702. There are two problems with relying on such anecdotal data, at least exclusively. First, anecdotal evidence is usually derived from insufficient sampling—a handful of select cases that do not reflect a cross section of the relevant population. Second, there is a strong possibility that anecdotal cases are not comparable to the facts of the case at bar. For example, if a number of people contract lung cancer through an alleged exposure to a toxic substance, a proper statistical study of the causative link between the toxic substance and the cancer will exclude the possibility of other sources for the injury (referred to as confounding factors), such as cigarette smoking. But a single case study, or other anecdotal evidence, is unlikely to be screened for confounding factors.

For example, in *Cavallo v. Star Enterprises*,[97] the plaintiff alleged that she suffered respiratory illness as a result of exposure to aviation jet fuel vapors that were released from the defendant's storage terminal. The plaintiff's expert toxicologist was prohibited from testifying after a *Daubert* hearing, and the court consequently granted summary judgment for the defendant. To support an opinion that the plaintiff's respiratory illness was caused by the jet fuel vapors, the toxicologist relied solely on case studies in which people who were exposed to the organic compounds in jet fuel suffered respiratory illnesses. The court held that reliance on these studies to form a conclusion was inconsistent with the scientific method. The court reasoned that "case reports are not reliable scientific evidence of causation, because they simply describe reported phenomena without comparison to the rate at which the phenomena occur in the general population or in a defined control group; do not isolate and exclude potentially alternative causes; and do not investigate or explain the mechanism of causation."[98]

This is not to say that case studies are completely irrelevant to an analysis consistent with the scientific method. The problem in *Cavallo* was that the case studies were, essentially, the *only* source upon which the expert based his opinion. In contrast, scientists often use case studies to spur further research or to confirm conclusions reached in more methodical studies.[99]

97. 892 F. Supp. 756, 767 (E.D. Va. 1995), *aff'd in pertinent part,* 100 F.3d 1150 (4th Cir. 1996).

98. *Id.*; *see also* Norris v. Baxter Healthcare Corp., 397 F.3d 878, 885 (10th Cir. 2005) (". . . case reports suffer from a similar failing. Case reports that state that some women with breast implants developed disease do not provide an adequate scientific basis from which to conclude that breast implants in fact cause disease. A correlation does not equal causation.").

99. *See, e.g.,* Wendell v. GlaxoSmithKline LLC, 858 F.3d 1227, 1236–37 (9th Cir. 2017) ("Although case studies alone generally do not prove causation, they 'may support other proof of causation.' Here, the experts relied not just on these studies—which not only examined reported cases but also used statistical analysis to come up with risk rates—but also on their own wealth of experience and additional literature.") (citation omitted); Cantrell v. GAF Corp., 999 F.2d 1007,

Reliance on Temporal Proximity

There are a number of cases in which a healthy plaintiff has become ill shortly after his exposure to a product. For example, in *Porter v. Whitehall Lab.,*[100] Porter came to the hospital to be treated for a fractured toe. He had no other documented health problems. He was given ibuprofen. Over the following month, the plaintiff took about thirty tablets. At the end of that month, the plaintiff was diagnosed with rapidly progressive glomerulonephritis (RPGN), a form of end-stage renal failure from which he would not recover. At the time, no studies demonstrated a link between ibuprofen use and RPGN.[101] Porter's experts, however, testified that ibuprofen caused his RPGN. The experts essentially based their opinions on Porter's prior good health and the temporal proximity between the ingestion of ibuprofen and his renal failure. The court concluded, however, that forming a conclusion on the basis of temporal proximity alone, in the absence of some established scientific connection between substance and illness, is inconsistent with the scientific method. By relying *solely* on temporal proximity, the expert fails to consider other possible explanations that a scientist must consider before drawing a conclusion. An expert who determines causation by relying only on temporal proximity commits the classic error of confusing correlation and causation.

But this is not to say that the temporal proximity between exposure and injury can never be relied upon in reaching an opinion on causation. The result in *Porter* might well have been different if published controlled studies and/or epidemiological evidence established some connection between his ibuprofen intake and RPGN. Then the temporal proximity between exposure and injury could be used as confirmatory data. This was the case in *Kennedy v. Collagen Corp.,* above, where the expert's opinion on causation was based, in part, on the plaintiff's having contracted lupus shortly after receiving collagen injections.[102] The expert could properly rely on this factor in light of the studies indicating a substantial connection between collagen and the autoimmune antibodies present

1014 (6th Cir. 1993) (expert testified that asbestos created a risk of laryngeal cancer, basing his conclusion on epidemiological evidence reported in the medical literature, and on the inordinately high incidence of persons at the plaintiffs' workplace whom the expert had personally diagnosed as having laryngeal cancer; the court held that the expert testimony was properly admitted, stating that "[n]othing in Rules 702 and 703 or in *Daubert* prohibits an expert from testifying to confirmatory data, gained through his own clinical experience, on the origin of a disease or the consequences of exposure to certain conditions").

100. 9 F.3d 607 (7th Cir. 1993).

101. There were studies suggesting a connection between ibuprofen and interstitial nephritis—a secondary kidney condition from which Mr. Porter also suffered. But there was no demonstrated connection between nephritis and RPGN. Nor did the studies showing a connection between nephritis and ibuprofen intake support causation as a result of the very modest dosage to which Porter was subjected.

102. 161 F.3d 1226, 1230 (9th Cir. 1998).

in lupus patients. As one court put it, "[a] temporal connection is entitled to greater weight when there is an established scientific connection between exposure and illness or other circumstantial evidence supporting the causal link."[103]

Insufficient Connection Between the Expert's Opinion and the Facts in the Case: The Requirement of "Fit"

Some experts may rely on a proper methodology, but then fail to connect the methodology with the facts of the case. This is the problem of "fit." An example of a failure to connect reliable expert testimony with the facts arose in *Harris v. Remington Arms Company, LLC*.[104] In *Harris*, the plaintiff brought a product liability action alleging that her Remington rifle had fired and injured her when the safety was turned off—but the trigger was not pulled. The plaintiff offered an expert witness to explain how the rifle could have fired without anyone pulling the trigger. The expert testified that the safety and the trigger mechanism bonded together after the plaintiff had engaged the safety and then stored the rifle in a cold room, causing a liquid bonding agent to solidify. According to the expert, therefore, disengaging the safety caused the trigger to be pulled simultaneously. There was no challenge to the methodology relied upon by the expert to conclude that the liquid bonding agent could solidify in cold temperatures. But this testimony did not fit the facts of the case because the plaintiff had undisputedly disengaged the safety at other times after taking the gun from the cold room, and it had not gone off. The court found that the trial court did not err in excluding the expert's testimony, because of its "misfit with the undisputed evidence."[105]

103. Curtis v. M&S Petroleum, Inc., 174 F.3d 661, 670 (5th Cir. 1999) (abuse of discretion to exclude the testimony of a doctor who concluded that the plaintiffs' health problems were caused by exposure to benzene; the expert relied not only on the "strong temporal connection between the refinery workers' exposure to benzene and the onset of their symptoms" but also on extensive scientific literature establishing a connection between benzene and the symptoms suffered by the plaintiffs). *See also* Claussen v. M/V New Carissa, 339 F.3d 1049 (9th Cir. 2003) (in concluding that damage to oyster beds was caused by an oil spill, the expert could rely on the temporal proximity between the spill and the damage, when combined with field studies, government reports, and exclusion of other obvious causes, to reach a reliable opinion under *Daubert*).

104. 997 F.3d 1107, 1114 (10th Cir. 2021).

105. *Id.* (". . . the district court didn't reject Mr. Powell's methodology. The court instead concluded that Mr. Powell's opinion testimony didn't fit the undisputed evidence."); *see also* Bogosian v. Mercedes-Benz of N. Am., Inc., 104 F.3d 472, 479 (1st Cir. 1997). In that case, the plaintiff was run over by her car after putting it in park and exiting the vehicle. She sought to have an engineer testify about the phenomenon of "false park detent"—where the driver feels as if he or she has put the car in park, without looking at the gear shift, but the car is not actually in park. The court held that this testimony was properly excluded under *Daubert* because the testimony rested on a factual premise—that

Lack of Testing

One of the factors identified by *Daubert* for assessing the reliability of an expert's methodology pursuant to Rule 702 was whether the expert's technique or theory can be or has been tested. Courts sometimes exclude experts when they have not tested theories or techniques that are capable of being tested. For example, in *Truck Insurance Exchange v. MagneTek Inc.*,[106] the Tenth Circuit held that the trial court properly excluded plaintiff's expert testimony concerning the novel theory of "pyrolysis." This theory hypothesized that wood could ignite at temperatures much lower than normal under particular circumstances. The expert testimony was excluded because neither the experts nor others in the scientific community had tested the novel theory sufficiently to demonstrate its scientific reliability. The Tenth Circuit affirmed, explaining that the importance of testing as a factor in determining reliability is at its highest when an expert proposes a theory that modifies otherwise well-established knowledge about regularly occurring phenomenon—such as the temperature at which wood will ignite.[107]

The problem of lack of testing also appears to arise frequently in product liability cases, where the plaintiff calls an expert to testify that the defendant should have designed a product differently. If the alternative design proposed by the expert has never been made and tested, either by the expert or anyone else, courts are likely to prohibit the expert from testifying that such a design was a viable and safer alternative. As the Seventh Circuit noted in *Cummins v. Lyle Indus.*,[108] a number of factors must go into a reliable conclusion that an alternative design should have been employed, such as the compatibility of the design with existing systems, relative efficiency, maintenance costs, ease of servicing, and the effects of price. As the *Cummins* court put it: "Many of these considerations are product-and-manufacturer-specific, and most cannot be determined reliably without testing."[109]

the plaintiff did not look at the console shift before turning off the car—that was at odds with the plaintiff's own testimony: "The district court appropriately found it very odd that Bogosian would present an expert witness who would testify that her own unwavering testimony was incorrect."

106. 360 F.3d 1206, 1211–12 (10th Cir. 2004).

107. *See also* United States v. Gabaldon, 389 F.3d 1090, 1099 (10th Cir. 2004) (testimony of an accident reconstructionist in a murder case that the internal geometry of a car would have made it difficult for the defendant to reach the victim with enough power to inflict blows was properly excluded where he conceded that his theory could have been tested by placing defendant in the car and taking measurements, but that no such testing was done). *But see* Bitler v. A.O. Smith Corp., 400 F.3d 1227, 1235 (10th Cir. 2004) (expert opinion as to cause of explosion was properly permitted; the fact that the expert did not test his theory about the cause of the gas leak at issue was not dispositive where his opinion was based upon "known science of copper sulfide particulate contamination as a cause of propane gas leaks").

108. 93 F.3d 362, 369 (7th Cir. 1996).

109. *Id.* at 370–71 (affirming exclusion of alternative design testimony because the alternative design had not been tested, and the expert's opinions lent themselves to testing and substantiation by the scientific method).

But the *Daubert* factors are flexible, and Rule 702 is context-specific. There is no absolute requirement that a conclusion or theory be tested to be admissible. In the specific area of design safety, the feasibility of testing and the expense of litigation must be taken into account. For example, a design engineer may seek to testify that a car could have been designed more safely. A court should not exclude the testimony solely because the expert has not built a car employing the alternative design. A good example of the proper approach to testing after *Daubert* is found in *Tassin v. Sears, Roebuck & Co.*[110] Tassin was injured while operating a power saw and proffered an engineer who concluded that alternative designs were safer and that the defendant failed to provide adequate warnings. The district court analyzed the admissibility of this testimony as follows:

> It may well be that an engineer is able to demonstrate the reliability of an alternative design without conducting scientific tests, for example, if he can point to another type of investigation or analysis that substantiates his conclusions. For example, an expert might rely upon a review of experimental, statistical, or other technical industry data, or on relevant safety studies, products, surveys, or applicable industry standards. He could also combine any one or more of these methods with his own evaluation and inspection of the product based on experience and training in working with the type of product at issue. The expert's opinion must, however, rest on more than speculation, he must use the types of information, analyses and methods relied on by experts in his field, and the information that he gathers and the methodology that he uses must reasonably support his conclusions. If the expert's opinions are based on facts, a reasonable investigation, and traditional technical/mechanical expertise, and he provides a reasonable link between the information and procedures he uses and the conclusions he reaches, then [the opinion may be admitted despite a lack of testing].[111]

Applying these standards to the facts, the *Tassin* court excluded the expert's testimony on certain alternative designs based on parts that he had never tested or even seen, and the safety of which was not supported by the tests of others or by any relevant literature. However, the court held testimony as to another alternative design admissible, where the expert had actually conducted some testing and where the safety of the product received support from the relevant literature. After finding sufficient reliable support to admit this opinion under Rule 702, the court explained that the expert's failure to test more systematically and extensively presented a question of the weight to be afforded to the opinion. Finally, the court held the expert's conclusion as to inadequate warnings to be admissible. The alternative warnings suggested by the expert had not been scientifically tested. But the court found that testing as to warnings (as opposed to testing alternative designs) was not critical where the expert had

110. 946 F. Supp. 1241, 1247–48 (M.D. La. 1996).
111. *Id.*

substantial experience in both product design and in preparing product manuals and warnings.

The Problem of Overstatement

Some expert witnesses may utilize a methodology that is sufficiently reliable and that may be applied helpfully to the facts of a particular case but may err in their application by overstating the findings that the methodology is capable of producing (particularly during trial testimony). Overstatement is especially endemic in the testimony of forensic experts. Over the past two decades, various scientific bodies have raised serious questions about the reliability of forensic techniques that had theretofore been admitted without much controversy. The two most important reports are:

1. National Research Council of the National Academy of Sciences (NAS), Strengthening Forensic Science in the United States: A Path Forward (2009). This detailed report essentially concludes that with one exception, none of the forensic techniques currently employed—such as ballistics, handwriting, and hair identification—meet the standards of science. The exception is identification of DNA from a single source.
2. The President's Council of Advisors on Science and Technology (PCAST), Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016). This report provides an exhaustive analysis of why forensic feature comparison methods (specifically ballistics, handwriting, fingerprinting, footwear, hair, and DNA analysis from complex mixtures) have not been validated, and how at least some of them can be strengthened so that they have validity. Particular attention is given to the problem of experts overstating their results—e.g., making statements that fingerprint identification has a "zero rate of error" when in fact the process is based on subjective judgment that by definition carries a rate of error.

Despite these important reports, most courts have admitted forensic expert opinions as reliable technical testimony under *Kumho* even though they are not based on the scientific method.[112] Some courts, though, have rightly focused on

112. *See, e.g.,* United States v. Baines, 573 F.3d 979, 989 (10th Cir. 2009) (conceding that there are "multiple questions regarding whether fingerprint analysis can be considered truly scientific in an intellectual, abstract sense" but concluding that "nothing in the controlling legal authority we are bound to apply demands such an extremely high degree of intellectual purity" and that "fingerprint analysis is best described as an area of technical rather than scientific knowledge"); United States v. Johnson, 2019 U.S. Dist. LEXIS 39590 (S.D.N.Y.) (finding ballistics identification to be admissible, and claiming that the NAS and PCAST reports impose "demanding scientific standards" that "require a level of certainty and infallibility not properly applied in a courtroom").

preventing forensic experts from *overstating* their results—for example, prohibiting such experts from testifying that they are "certain" that there is a "match," or that the rate of error is "zero" or "infinitesimal." The leading case on preventing overstatement is *United States v. Glynn*,[113] in which the district court found that because ballistics identification is based on subjective judgment, the expert could not testify that he was a "scientist" or that the rate of error for his identification was zero. The court allowed the expert to testify but held that he could state only that it was "more likely than not" that the bullet fragment came from the defendant's gun—as that was all the certainty that the subjective methodology could support. The trial judge explained that it is particularly important for judges to regulate expert overstatement as part of the gatekeeping role because:

> [O]nce expert testimony is admitted into evidence, juries are required to evaluate the expert's testimony and decide what weight to accord it, but are necessarily handicapped in doing so by their own lack of expertise. There is therefore a special need in such circumstances for the Court, if it admits such testimony at all, to limit the degree of confidence which the expert is reasonably permitted to espouse.[114]

While some other courts have followed suit and have sought to control overstatement in expert testimony,[115] many others have not.[116] Because the problem of overstatement has not been well regulated by the courts, the Advisory Committee on Evidence Rules, starting in 2017, began considering whether Rule 702

113. 578 F. Supp. 2d 567, 575 (S.D.N.Y. 2008).

114. *Id.*

115. *See, e.g.,* United States v. Machado-Erazo, 901 F.3d 326 (D.C. Cir. 2018) (error for a cell phone location expert to testify to the precise location of the defendant's phone, where the current technology can locate a phone only within a general area); United States v. Parker, 871 F.3d 590 (8th Cir. 2017) (trial court prohibited the expert from testifying that she was "100% sure" or "certain" that the relevant guns matched the relevant shell casings); United States v. Hill, 818 F.3d 289, 295 (7th Cir. 2016) (declaring that cell site analysis expert testimony should include a "disclaimer" regarding accuracy; the expert should not "overpromise on the technique's precision or fail to account for its flaws"; cell site testimony was admissible where the agent testified that the defendant's phone records were "consistent" with him being at or near relevant locations at relevant times, but clarified that he could not state whether a phone was "absolutely at a specific address"); United States v. White, 2018 U.S. Dist. LEXIS 163258 (S.D.N.Y.) (finding proposed ballistics testimony was admissible, subject to the limitation that the expert could not testify to any specific degree of certainty that there was a ballistics match between the firearms seized from the defendant and those used in the various shooting incidents).

116. *See, e.g.*, United States v. Straker, 800 F.3d 570 (D.C. Cir. 2015) (fingerprint expert allowed to testify that there was a "zero rate of error in the methodology"); United States v. Hunt, 2020 WL 2842844 (W.D. Okla. 2020) (the defendant asked the court "to place limitations on the Government's firearm toolmark experts because the jury will be unduly swayed by the experts if not made aware of the limitations on their methodology"; the court allowed the expert to testify to a "reasonable degree of certainty" even though the Department of Justice standards do not permit its experts to testify in those terms).

should be amended to prohibit overstatement by experts. After careful consideration, the committee concluded that overstatement was already within the court's gatekeeping responsibility under Rule 702, as amended in 2000, because an expert who overstates a conclusion is not reliably applying her basis and methodology. But the committee determined that the obligation to regulate overstatement should be highlighted for the courts with a modest amendment to Rule 702(d). The 2023 amendment to Rule 702(d) discussed above requires the court to find that "the expert's opinion reflects a reliable application of the principles and methods to the facts of the case."[117] The amendment instructs trial courts to carefully evaluate the actual opinion rendered—the opinion itself and not just the process of applying a methodology—to make sure that it does not go beyond what the expert can reliably conclude. The Committee Note to the 2023 amendment emphasizes that the focus of the amendment is on preventing experts from overstating their opinions:

> Rule 702(d) has . . . been amended to emphasize that each expert opinion must stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology. Judicial gatekeeping is essential because just as jurors may be unable, due to lack of specialized knowledge, to evaluate meaningfully the reliability of scientific and other methods underlying expert opinion, jurors may also lack the specialized knowledge to determine whether the conclusions of an expert go beyond what the expert's basis and methodology may reliably support.

While the Committee Note voices objection to testimony about a "zero rate of error," the intent of the amendment is to require court scrutiny of any attempt by a forensic expert to testify to greater certainty than is supported by the methodology. A common and problematic form of expert overstatement occurs when an expert witness purports to testify "to a reasonable degree of scientific certainty." Expert overstatement was a significant focus of the PCAST report on forensic feature-comparison methods.[118] In particular, the PCAST report recommended that courts "should never permit scientifically indefensible claims, such as . . . proof to a reasonable degree of scientific certainty."[119] A report from the National Commission on Forensic Sciences similarly decried this type of expert overstatement and proposed that courts should forbid experts from stating their conclusions to a "reasonable degree of [field of expertise] certainty," because that term has no scientific foundation. The Department of Justice (DOJ) has adopted testimonial protocols that prohibit the use of the "reasonable degree of certainty" language for certain feature-comparison experts, and that impose important limitations on testimony regarding rates of error. For example, a DOJ

117. The amendment took effect on December 1, 2023.

118. President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. (Sept. 20, 2016).

119. *Id*. at 145.

directive regarding toolmark testimony provides, in pertinent part: "An examiner shall not use the expressions 'reasonable degree of scientific certainty,' 'reasonable scientific certainty,' or similar assertions of reasonable certainty in either reports or testimony, unless required to do so by a judge or applicable law."[120] It is for this reason that the Advisory Committee Note to the 2023 amendment to Rule 702 specifically explains: "Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error." Courts should thus prohibit, where possible, the use of problematic testimony to a "reasonable degree of scientific certainty" notwithstanding significant past federal precedent in which such testimony was permitted or even required.[121]

Experts Relying on Technical or Other Specialized Knowledge

As discussed above, the Supreme Court's *Kumho* decision clarified that Rule 702 and the *Daubert* reliability standard apply to expert opinion testimony based upon technical or other specialized knowledge, as well as to scientific testimony. Even though the gatekeeper function applies to nonscientific expert testimony, it stands to reason (as the Court in *Kumho* recognized) that the specific *Daubert* factors will often have to be rejected or modified when reviewing most nonscientific expert testimony. A court should not exclude an expert on automobile repair on the ground that he has not published his research in a peer-reviewed journal. A court should not reject the testimony of a sociologist simply because she cannot be definitive about a possible rate of error in her findings.

120. United States Department of Justice Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Fracture Examination 3–4, https://perma.cc/P6D4-BCC9.

121. There can be a complication in rejecting the reasonable degree of certainty standard in civil cases, however. Some states appear to require a reasonable certainty standard as a matter of state substantive law—which is controlling in diversity cases, assuming that in fact it is substantive. *See, e.g.,* Miranda v. County of Lake, 900 F.3d 335 (7th Cir. 2018) ("In Illinois, proximate cause must be established by expert testimony to a reasonable degree of medical certainty."); Day v. United States, 865 F.3d 1082 (8th Cir. 2017) (under Arkansas law, a medical expert must testify that "the damages would not have occurred" without the defendant's negligence; expert's opinion "must be stated within a reasonable degree of medical certainty or probability"). For this reason, the advisory committee's note to the 2023 amendment to Rule 702 explains that the "amendment does not, however, bar testimony that comports with substantive law requiring opinions to a particular degree of certainty."

An example of this necessary flexibility is found in *Tyus v. Urban Search Management*.[122] The plaintiffs in *Tyus* alleged that advertising for a rental building violated the Fair Housing Act because it targeted only white lessees. The plaintiffs proffered a social science expert who would have testified to how an all-white advertising campaign affects African Americans. The court declared that the central teaching of *Daubert*—that expert testimony "must be tested to be sure that the person possesses genuine expertise in a field and that her court testimony adheres to the same standards of intellectual rigor that are demanded in her professional work"—was fully applicable to the testimony of experts in the social sciences. However, recognizing the need for a flexible application of *Daubert*, the court noted the following caveat:

> It is true, of course, that the measure of intellectual rigor will vary by the field of expertise and the way of demonstrating expertise will also vary. Furthermore, we agree . . . that genuine expertise may be based on experience or training. In all cases, however, the district court must ensure that it is dealing with an expert, not just a hired gun.

The *Tyus* court found that the trial court erred in excluding the expert, because his research was based on peer-reviewed articles, and his "focus group" method was a well-accepted methodology in the field of social science. In other words, the expert brought the same intellectual rigor to his courtroom testimony as he employed in his life as a social scientist.

Where expert testimony is not adaptable to being put through the wringer of falsifiability, publication, and peer review, the trial judge still has the obligation to determine whether the testimony is properly grounded, well reasoned, and not speculative. If there is a well-accepted body of learning, experience, and basic assumptions in the field, then the expert's testimony should be grounded in that learning and experience to be reliable. The more subjective and controversial the expert's inquiry, the more likely the testimony is to be excluded as unreliable. The opinion in *Frymire-Brinati v. KPMG Peat Marwick*[123] is instructive. In this action for securities fraud, the plaintiff's expert, an accountant, was allowed to testify that a Peat Marwick audit had improperly certified that property interests had a certain value when in fact they were worth much less. To reach this conclusion, the accountant used a discounted cash flow analysis, by which he assessed value solely on the basis of net, rather than potential, cash flow. He did not explain why he failed to consider potential cash flow, even though standard principles of valuation find it to be an important factor. Reversing a judgment for the plaintiff, the court held that the trial court abused its discretion in admitting the expert's valuation, because the expert's

122. 102 F.3d 256, 263 (7th Cir. 1996).
123. 2 F.3d 183, 188 (7th Cir. 1993).

methodology was faulty: by failing to consider potential cash flow, the expert's methodology would lead to the conclusion that "raw land is worthless and that a large office building in the final stages of construction also has no value even though it is fully leased out and could be sold for a hundred million dollars." The Seventh Circuit held that the expert's testimony lacked validity under *Daubert*, where his litigation methodology would not have been acceptable for clients outside the courtroom.

A particular problem after *Kumho* involves the nonscientific expert who testifies solely on the basis of experience. The example from *Kumho* is the perfume tester who sniffs a substance and identifies its brand and provenance. When asked how he can make such a specific identification, he responds, "I know it when I smell it and I smell it a lot." Should he be permitted to testify solely on this conclusory experience-based analysis? Put another way, must the trial court take the expert's word for it? The Court in *Kumho* implies that it is inconsistent with the gatekeeper function to allow an experience-based expert to testify without any explication of *how* he or she comes to a conclusion and *why* he or she believes it is correct. This implication—that experience-based experts must justify their conclusions as reliable—was codified in the 2000 amendment to Rule 702.

For example, in *United States v. Rodriguez*,[124] law enforcement officers testified as experts interpreting intercepted phone calls. But the officers testified about uncommon terms or phrases that they encountered for the first time in the investigation. They relied on experience, but the court found error in admitting their testimony, because "the officers generally did not offer an explanation for how they arrived at their interpretations, nor did the court require them to provide one." Any explanations the officers did offer for arriving at their interpretations "failed to evince indicia of reliability or methodological rigor." The court noted that "an officer's qualifications, including his experience with investigations and intercepted communications, are relevant but not alone sufficient to satisfy Federal Rule of Evidence 702" because "Rule 702 requires district courts to assure that an expert's methods for interpreting the new terminology are both reliable and adequately explained."[125] To be sure, law enforcement officers are frequently permitted to give expert testimony regarding the meaning of code words, particularly in gang- and drug-related cases, so long as they offer the requisite explanation of their methodology.[126]

124. 971 F.3d 1005, 1018 (9th Cir. 2020).

125. *Id.*; *see also* United States v. Hermanek, 289 F.3d 1076 (9th Cir. 2007) (expert testimony translating purported coded conversation should have been excluded; although the government made an adequate showing of the witness's experience, the witness did not explain in detail the methods he used to arrive at his interpretation of words that he was not familiar with before the investigation).

126. *See* United States v. Wilson, 484 F.3d 267, 275 (4th Cir. 2007) (explaining that "courts of appeals have routinely held that law enforcement officers with extensive drug experience are

Conclusion

Exercising the gatekeeping function under *Daubert* and Rule 702 is sometimes daunting and often requires a substantial expenditure of judicial resources. But the alternative—leaving questions of expert reliability to the jury—is inconsistent with Rule 702 and established Supreme Court precedent. Thus, the trend over time has been for judges to exercise closer scrutiny of expert testimony, and this trend may be expected to continue as evidence in litigation becomes more complex.

qualified to give expert testimony on the meaning of drug-related code words" and affirming admission of such testimony where expert carefully "summarized his methodology").

How Science Works

MICHAEL WEISBERG AND ANASTASIA THANUKOS

Michael Weisberg, Ph.D., is Bess W. Heyman President's Distinguished Professor of Philosophy and Deputy Director of Perry World House, University of Pennsylvania.

Anastasia Thanukos, Ph.D., is principal editor of the University of California Museum of Paleontology, University of California, Berkeley, and editor of the Understanding Science project.

CONTENTS

FIGURES

TABLE

Introduction

Science typically appears in the courtroom as scientific evidence presented by witnesses deemed experts by the courts. Such evidence may be considered by a jury or judge, depending on the type of case. Judges are the gatekeepers for this evidence, determining what information and testimony will be allowed to enter the court and influence the outcome of a trial, a challenging and essential role for several reasons. The broad public, from which juries are selected, has a limited understanding of many basic, well-established, and uncontroversial scientific concepts.[1] Yet in a trial, a jury may have to weigh evidence regarding advanced scientific knowledge that is still emerging and about which lines of evidence may conflict or be inconclusive. Complicating matters, public relations campaigns have misled the public about the true state of scientific consensus regarding certain scientific issues.[2] Further, scientific expertise and the trappings of science can be made to seem persuasive, even when there is little substance to them, making it more difficult for a jury to come to a just decision when presented with shaky evidence about which they have limited scientific understanding. Thus, the responsibility to ensure that the testimony the jury hears is scientifically sound and meets minimal standards of reliability is an important one.

The purpose of this reference guide is to provide a nuts-and-bolts look at how science works today to frame subsequent reference guides of this reference manual and to inform judges' consideration of proffered scientific evidence and their decisions about what testimony to allow in a trial. Other reference guides will help judges assess questions of law (e.g., What opinions have shaped interpretation of the Federal Rules of Evidence?; see *The Admissibility of Expert Testimony*, in this manual) and questions specific to scientific disciplines (e.g., Does bitemark evidence emerge from "good" science and should it be permitted at trial?; see *Reference Guide on Forensic Feature Comparison Evidence*, in this manual). In this reference guide, we provide background and guidance to inform questions that turn on the processes and nature of science writ large: Does peer review ensure that a study is scientifically sound? What qualifies someone as an expert within the scientific community? Should the sponsor of scientific research (e.g., a government agency versus a company) factor into interpretation of its findings? Fairly evaluating such questions requires an understanding of the nature and process of science that goes beyond media and textbook portrayals. The view of science presented below is based on current research and thinking in an academic discipline that studies the theory and processes of scientific knowledge building: the philosophy of

1. Pew Research Center, What Americans Know About Science (March 2019).

2. Naomi Oreskes & Erik M. Conway, Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming (2010) [hereinafter Oreskes & Conway 2010].

science.[3] We focus in this reference guide on the Western conception of science as it currently operates, recognizing that science is a social and cultural practice that has changed and will continue to change over time.

What Is Science: "Textbook" Science vs. the Practice of Science

Science is both a body of knowledge and the process for building that knowledge based on evidence acquired through observation, experiment, and simulation. The term is accurately applied to knowledge on a wide variety of topics and to diverse lines of inquiry. These can include familiar disciplines in the natural sciences, like medicine and physics, as well as the social and behavioral sciences. Issues involving economics, history, education, politics, social trends,

Clarifying vocabulary: Scientific knowledge

Misconceptions about and alternative definitions of the terms *hypothesis, theory*, *model,* and *law* abound. Herein, we use the following definitions for these elements of scientific knowledge:

Hypothesis—a potential explanation for a natural phenomenon, regardless of the extent to which the explanation has been investigated or the amount of evidence supporting or refuting it

Theory—a concise, coherent, and predictive explanation for a broad range of natural phenomena that integrates and makes sense of many hypotheses

Model—a mathematical representation of hypothesized mechanisms and interactions within a system that can be used to investigate the impact of parameter changes on predicted system outcomes

Law—an often-mathematical statement of the relationship among observable phenomena

Note that within and outside of science, these terms are used inconsistently, and their popularity has risen and fallen in different disciplines and at different points in history. For further discussion of these terms, see sections titled "Not Necessarily Consensus" and "Myths About Science" below.

3. Readers interested in an accessible introduction to this discipline, its history, and current perspective on the nature and process of science are referred to Peter Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science* (2003) [hereinafter Godfrey-Smith 2003].

human opinions, and much more can be investigated with scientific rigor. This broad view of the sorts of topics within the purview of science aligns with the wide variety of topics addressed by later reference guides in this manual. In short, what makes a unit or body of knowledge scientific is the process by which it was established, not the topic it concerns.

What is this special process that makes an inquiry or investigation scientific? How do we generate scientific knowledge? Textbooks and media often present a sterile and simplified, but commonly held, view of the process of science, sometimes called the scientific method,[4] which goes something like this: First, a scientist forms a hypothesis, a potential explanation for a natural phenomenon. For instance, a scientist might hypothesize that a particular class of chemicals is causing a hole in the ozone layer. The scientist performs an experiment to test the idea, perhaps by releasing the chemical in an ozone-filled chamber and measuring changes in the gases present in the chamber. The scientist reaches a conclusion about the hypothesis—that it is correct because ozone decreased significantly in the presence of the chemical. The results of the experiment are then written up and published. Other scientists read the article and mainly agree with the conclusions, and, voila, a new scientific fact is established. This simplified view correctly calls out an essential feature of scientific knowledge building—that it is based on evidence from the natural world that can stand up to outside scrutiny—but is misleading in other ways.

Table 1. Textbook Science vs. Science in Practice

Common representation of the scientific method	Characteristics of science in practice
Scientific investigations follow a codified step-by-step method.	Scientists engage in many different activities in different orders as they pursue evidence and explanations.
Scientists work in isolation.	Most modern science is collaborative and depends on social interactions within the scientific community.
Scientists use experiments to gather evidence.	Scientists use different methods to gather evidence; an experiment is just one of these methods.
Scientific studies reach a firm conclusion.	Scientific conclusions are always revisable if warranted by the evidence.
A single convincing study leads to scientific consensus on a hypothesis.	Scientific consensus is generally reached over the course of multiple studies conducted by different groups pursuing different lines of inquiry.

4. William F. McComas, *The Principal Elements of the Nature of Science: Dispelling the Myths*, *in* The Nature of Science in Science Education: Rationales and Strategies 53 (1998) [hereinafter McComas 1998].

An Example of Science in Practice: Connecting CFCs to Ozone Depletion

Although scientific inquiry always involves a reliance on evidence and has certain other traits (see section titled "Key Traits of Science" below), there is no one scientific method.[5] For example, while scientists have reached consensus on the hypothesis that human production of a certain class of chemicals, chlorofluorocarbons (CFCs), depletes the ozone layer, the actual story is not a simple one. We shall tell this story here, and then refer to it, as well as to other scientific studies that have made their way into the courtroom, throughout this reference guide to illustrate key points about the modern practice of science.

In the 1970s, a medical researcher, curious about the source of the haze near his home, started trying to detect chemicals in the air that come exclusively from human activities—in this case, CFCs. He found them everywhere, even in Antarctica, far from where they were being produced.[6] The findings were presented at a conference, where they caught the attention of another scientist, Sherwood Rowland, who shared it with his postdoctoral researcher, Mario Molina. Molina then turned to the published literature to figure out what might happen to these molecules once released into the atmosphere.

Molina and Rowland published an article putting forth the hypothesis that CFCs deplete our protective ozone layer, but their evidence did not come from an experiment or even any new observations.[7] Instead, the pair brought together measurements collected by other researchers, calculations, and established knowledge about basic and atmospheric chemistry, arguing that if all these other ideas and estimates were true, a logical outcome is that CFCs pose a threat to the ozone layer. Later, experiments were performed that suggested the chemical reactions that Rowland and Molina reasoned *should* happen actually *did* happen. Other scientists incorporated these ideas into their models of the atmosphere and made predictions about what should be observed if Rowland and Molina's hypothesis were correct. Still other groups collected atmospheric evidence to test the predictions made by the models. Meanwhile,

5. Philip Kitcher, The Advancement of Science: Science Without Legend, Objectivity Without Illusions (1995) [hereinafter Kitcher 1995]; William C. Wimsatt, Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality (2007) [hereinafter Wimsatt 2007]. For an argument that *Daubert v. Merrell Dow Pharmaceuticals, Inc.* is partly founded on the myth of the scientific method, see Sheila Jasanoff, *Law's Knowledge: Science for Justice in Legal Settings*, 95 Am. J. of Pub. Health s49–s58 (2005), https://doi.org/10.2105/AJPH.2004.045732 [hereinafter Jasanoff 2005].

6. J. E. Lovelock, R. J. Maggs, & R. J. Wade, *Halogenated Hydrocarbons in and Over the Atlantic*, 241 Nature 194–96 (1973), https://doi.org/10.1038/241194a0 [hereinafter Lovelock et al. 1973].

7. Mario J. Molina & F. Sherwood Rowland, *Stratospheric Sink for Chlorofluoromethanes: Chlorine Atom-Catalyzed Destruction of Ozone*, 249 Nature 810–12 (1974), https://doi.org/10.1038/249810a0 [hereinafter Molina & Rowland 1974].

the CFC industry backed another scientist to oppose the hypothesis, and Rowland and Molina checked some of the old measurements on which they had based their hypothesis and found them to be inaccurate. After correcting those numbers, models were updated, compared, and updated again. More data were collected and eventually, eight years after the hypothesis was first published, researchers discovered a thinning of the ozone layer over Antarctica much more extreme than expected.[8] This led to more revisions of the hypothesis, a few additional twists and turns, and eventually a Nobel Prize, the Montreal Protocol, which phased out CFC production, and today, a recovering ozone layer. Concurrently, ozone depletion has made its way into the courts as established science. For example, the National Resource Defense Council (NRDC) brought suit against the Environmental Protection Agency (EPA) for making decisions that violated the Montreal Protocol. The NRDC was found to have standing because of the increased risk of skin cancer that NRDC members would experience as a result of ozone destruction.[9] Ozone depletion was treated as a fact in that case, reflecting the scientific consensus on the issue.

In this case, science worked exactly as it ought. Hypotheses were sifted through, scrutinized by different parties with different interests, compared against multiple lines of evidence, and iteratively modified, ultimately leading to reliable and actionable scientific knowledge. Such examples are typical of science as it is actually practiced: Rarely do scientists apply a simple linear recipe, rely on a single critical experiment, or reach an immediate and firm conclusion. While the complex and iterative processes that went into establishing depletion of the ozone layer by CFCs are commonplace in science, the speed with which societal and political action followed scientific consensus in this case may be unusual.[10] We also note that this example of scientific consensus building played out over many years, and that, for good reason, much of the scientific testimony presented at trial will concern hypotheses with a much lower degree of certainty and that the scientific community has had less time to scrutinize. Nevertheless, throughout this reference guide the CFC/ozone example will serve as a valuable reminder of how science generally works, as well as the power of science to spur real-world change—a function that is often mediated by the courts and so further highlights the responsibility of the judge in evaluating the reliability of scientific testimony and in appointing scientific experts as appropriate.

8. J.C. Farman, B.G. Gardiner, & J.D. Shanklin, *Large Losses of Total Ozone in Antarctica Reveal Seasonal* ClO_x/NO_x *Interaction*, 315 Nature 207–10 (1985), https://dx.doi.org/10.1038/315207a0 [hereinafter Farman et al. 1985].

9. Nat'l Rsch. Def. Council v. Env't Prot. Agency, 464 F.3d 1 (D.C. Cir. 2006).

10. Oreskes & Conway 2010, *supra* note 2.

The Science Flowchart

Science is complex, iterative, dynamic, and social. It emerges out of the work of fallible people, shaped by their unique backgrounds, perspectives, and motivations, engaged in different activities in different orders as they come up with new explanations, and test and retest those ideas.[11] Through this process, biases are identified and corrected, inaccurate ideas are eventually rejected, and discrepancies are negotiated and resolved. This practice does lead to scientific consensus, but over many years, investigations, and interactions, not in the context of a single study. These key points are illustrated by the science flowchart (Figure 1), which more accurately represents the process of science than does the five-step scientific method frequently found in textbooks.[12]

This flowchart can help us frame scientific inquiry into a wide range of topics. It applies to the natural sciences, as well as social sciences and engineering research. It applies to both qualitative and quantitative research conducted in laboratories, in the field, in clinical settings, and through the examination of historical records and other datasets. It applies to research conducted by academics, engineers, government workers, industry employees, and experts hired by litigation teams. Any line of inquiry that aims to establish objective, reliable knowledge based on evidence can be examined using this lens. While such knowledge-building activities do not follow a prescribed step-by-step process, they *do* have some traits in common. At some point, in some way, someone gets an idea (often called a hypothesis) about how to solve a problem, explain a phenomenon, or answer a question; at some point, in some way, that hypothesis is tested against evidence; at some point, in some way, the evidence and hypothesis are scrutinized by others with relevant expertise; and if the hypothesis holds up to this scrutiny through multiple rounds of testing, people will use it—to solve problems, make decisions, and/or build more knowledge.

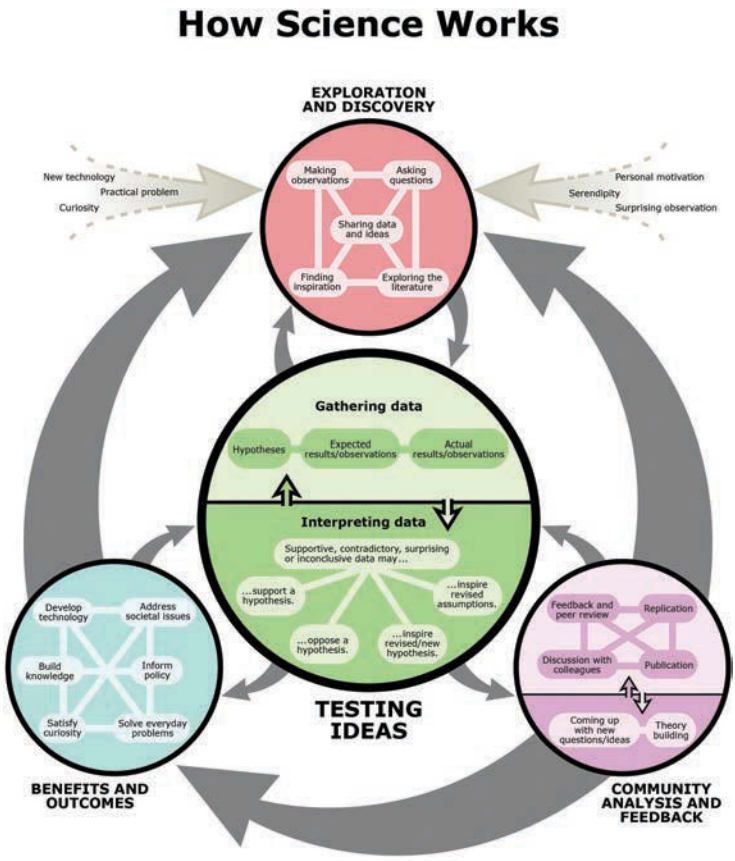
If mapped onto the science flowchart, our example of the discovery of the link between CFC production and the hole in the ozone layer would produce a winding path that circles back on itself, proceeding through the testing phase multiple times. A map of the investigations establishing that cigarettes cause lung cancer would follow a different path, as would a map of social science and psychological research examining factors that impact eyewitness recall.[13] When a judge must evaluate

11. For example, see David L. Hull, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science* (1988) [hereinafter Hull 1988]; Kitcher 1995, *supra* note 5; Helen E. Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (1990), https://doi.org/10.2307/j.ctvx5wbfz [hereinafter Longino 1990]; & Wimsatt 2007, *supra* note 5.

12. University of California Museum of Paleontology, Understanding Science (2022), https://perma.cc/MAB9-NYWL [hereinafter UCMP 2022].

13. Each of these lines of research has, at some point, been applied in the courtroom. *See* Nat'l Rsch. Def. Council v. Env't Prot. Agency, 464 F.3d 1 (D.C. Cir. 2006); United States v. Philip Morris USA Inc., 566 F.3d 1095 (D.C. Cir. 2009); *and* Sanders v. City of Chi. Heights, No. 13 C

Figure 1. Science is complex, iterative, dynamic, and social. It does not proceed according to a linear, step-by-step process.



Source: © UC Museum of Paleontology Understanding Science, www.understandingscience.org.

whether expert scientific testimony is scientifically sound, it is not a simple matter of ensuring that a particular five-step process was followed. Instead, the judge's task requires a deeper examination of the available evidence and methods by which it was arrived at, as well as an assessment of how the community of experts in this area has evaluated or would evaluate the evidence and reasoning in question.

Widely accepted scientific knowledge is built over the course of many iterations through this flowchart—a process that often takes years. For simplicity's

0221 (N.D. Ill. Aug. 18, 2016). For more on eyewitness identification, see Thomas D. Albright & Brandon L. Garrett, *Reference Guide on Eyewitness Identification*, in this manual.

sake, judges might wish to be presented only with scientific testimony based on mature research programs, in which many studies by different groups have been conducted, evidence has been scrutinized from different viewpoints, and some level of consensus has been reached; but, of course, disputed facts in litigation may hinge on questions of science that are still emerging, and so evidence is lacking, or that are so specific that new research must be conducted to inform the litigation. Such testimony may necessarily rely on one or just a few tests that have not yet received broad scrutiny by the relevant scientific community, and it may be that the scientific investigations that would best inform litigation have not been performed; neither situation renders inadmissible testimony based on the well-conducted and informative (though not conclusive) studies that *have* been performed. Despite our reliance on an example in which scientific consensus was eventually achieved, we emphasize that scientific consensus is not a requirement for testimony to be admissible, as described in the section below and in *The Admissibility of Expert Testimony* in this manual. If two experts present conflicting scientific testimony, and the testimony for each expert is grounded in proper scientific methodology, the court may admit the testimony of both experts and allow the trier of fact (i.e., jury or judge) to resolve the conflict.

This reference guide examines the progression of science from speculation to wide acceptance, as illustrated by the CFC/ozone example, to assist with judges' evaluation of the admissibility of scientific evidence and develop their understanding of the epistemology of science.[14] Subsequent reference guides will discuss the strengths and weaknesses of individual studies and lines of investigation in various scientific disciplines relevant to the judiciary.

The Science Flowchart and Admissibility

Federal Rule of Evidence 702 outlines the findings a judge must make to admit expert testimony over an objection.[15] According to Rule 702, such testimony is

14. Jasanoff 2005, *supra* note 5, laments the epistemological sophistication *Daubert* would seem to require to recognize "good science," as well as judges' lack of training in this area—a challenge that we hope this manual will help remedy.

15. While the testimony of expert witnesses identified by the parties is the most common way that science makes its way into the courts, it is not the only way. Court-appointed experts may perform a variety of functions and services. They may serve as witnesses themselves under Federal Rule of Evidence 706; advise the court on assessment of evidence that might, for example, be relevant for admissibility decisions; help mediate settlements; and/or provide background information to the judge or jury. In litigation that involves scientific evidence, the use of such independent scientific experts may help address potential challenges to reaching a just resolution that stem from aspects of the process of science and the process of litigation highlighted herein. These include the complexities of the process of science outlined above; the importance of community analysis and

admissible when the judge finds that (1) the witness is qualified to offer the testimony, (2) the witness's expertise will help determine questions of fact, (3) the testimony is based on sufficient facts or data, (4) the testimony results from reliable principles and methods, and (5) the witness has reliably applied those principles and methods to the facts of the case. In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the Supreme Court found that expert scientific testimony need not be "generally accepted" to be admitted under Rule 702; instead, the Court instructed federal judges to exercise a "gatekeeping" role regarding scientific evidence, admitting for consideration only such evidence that is grounded in an appropriate "scientific methodology."[16] The Court indicated that an appropriate scientific methodology often includes the process of formulating hypotheses and conducting experiments to test the validity of the hypotheses. The Court then provided a set of illustrative criteria suggestive of the sorts of considerations judges must make in determining whether the proposed testimony meets the standard of an appropriate scientific methodology.[17] The factors selected as examples by the Court include:

1. whether it can be and has been tested;
2. whether it has been subjected to peer review and publication;
3. whether it has a known error rate;
4. the existence and maintenance of standards and controls; and
5. whether the theory or technique employed by the expert is generally accepted in the scientific community.

The Ninth Circuit clarified another relevant consideration relating to *Daubert*: whether the research was conducted independently of the particular litigation or dependent on an intention to provide the proposed testimony. Subsequent

feedback in sorting through scientific explanations as described in the section titled "Science as a Human and Community Endeavor" below; the time all this takes to play out; the wide variety of valid scientific methodologies (described in the section titled "Scientific Methodologies and Data" below); the fact that individual scientists are human and inherently biased (described in the section titled "Bias" below); the depth of the subject matter knowledge that may be required to understand and evaluate scientific evidence; and of course the stakes that each party has in the testimony of their proffered expert witnesses and potential biases introduced in selecting those witnesses. Resources are available to assist judges in this process of identifying and securing a court-appointed expert—for example, the CASE program of the American Association for the Advancement of Science. For more on the services performed by court-appointed experts, see American Association for the Advancement of Science, *Court Appointed Scientific Experts: A Handbook for Experts Version 3.0* (2002), https://perma.cc/5FHD-GKQL, and Daniel L. Rubinfeld & Joe S. Cecil, *Scientists as Experts Serving the Court*, 147 Daedalus 152–63 (2018), https://doi.org/10.1162/daed_a_00526.

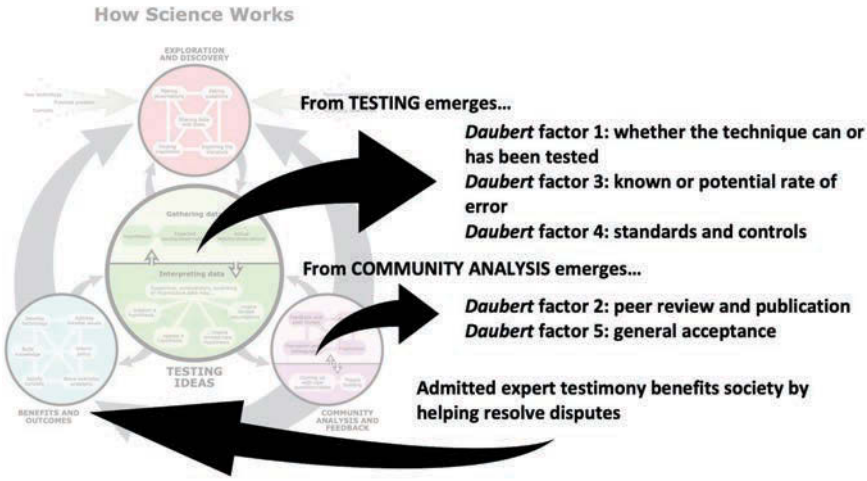
16. 509 U.S. 579 (1993).

17. For more on *Daubert* and admissibility, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

decisions by the Supreme Court clarified the standards for scientific and technical evidence.[18]

These legal standards for admissibility under *Daubert* can be mapped onto the elements of the flowchart, as illustrated by Figure 2. The central region of the flowchart outlines the logic of testing hypotheses against evidence, the first of the *Daubert* factors outlined above. Testing is essential to both the process of science and *Daubert*. If a hypothesis or testimony is not supported by or derived from evidence in some form, it cannot be based on scientific methodologies. The process of testing leads to two more of *Daubert*'s illustrative factors. The known or potential rate of error for scientific techniques or processes (the third *Daubert* factor) can be derived statistically from sufficient rounds of testing and data collection. The appropriate standards and controls required by a technique or process (the fourth *Daubert* factor) are likewise established in this way. Error rates, standards, and controls most clearly impact the consideration of studies that involve analytic and assessment techniques, such as sobriety tests, forensic identification techniques, and DNA analysis.

Figure 2. Relationships among the *Daubert* factors and the process of science.



Source: Image reproduced and adapted under an Attribution-NonCommercial-ShareAlike 4.0 (BY-NC-SA 4.0) Creative Commons License (see https://creativecommons.org/licenses/by-nc-sa/4.0/). Original image © UC Museum of Paleontology Understanding Science, www.understandingscience.org.

18. These standards were then codified through amendments to Federal Rule of Evidence 702. For more on Federal Rule of Evidence 702, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

The flowchart's *Community analysis and feedback* area highlights peer review and publication (the second *Daubert* factor) as an important element of scientific knowledge building. For many areas of science, peer-reviewed journal articles are the usual means through which the expert scientific community assesses hypotheses and evidence (see section titled "Peer Review" below). However, other scenarios are also possible. For example, an engineering technique or piece of computer code might be scrutinized through direct use by the relevant community to see if it performs as expected. Research conducted solely for litigation may be scrutinized by the judge, other expert witnesses, and a jury.[19] Finally, it is through interactions within *Community analysis and feedback* and multiple iterations through the *Testing ideas* region that hypotheses, theories, and techniques gain widespread acceptance (the final *Daubert* factor; see section titled "Achieving Scientific Consensus" below). And of course, the final result of the judge's decisions on the admissibility of scientific evidence represents one of the ways that benefits (the lower left area of the flowchart) are derived from the process of science, by helping resolve disputes through litigation.

Textbook portrayals of a five-step scientific method evoke a scientist in an academic institution driven by curiosity alone.[20] In contrast, the upper area of the science flowchart and the *Benefits and outcomes* region illustrate the variety of factors that motivate scientific investigations. Engineering research may be driven by the need to solve a particular problem, and as *Kumho Tire Co. v. Carmichael* held, the provenance of such specialized knowledge can be meaningfully compared to standards and methodologies of good science.[21] Career advancement, prestige, and funding are likewise incentives for many scientists.[22] And as described in *The Admissibility of Expert Testimony*, in this manual, scientific research may also be conducted for the sole purpose of testifying at trial.

The fact that a study was undertaken in the service of litigation does not itself make the research unscientific or unreliable.[23] *All* scientists have some sort of motivation for their work, and this does not preclude scientific knowledge

19. Jasanoff 2005, *supra* note 5, describes this role of the judiciary in helping produce certain sorts of scientific knowledge, acknowledging the non-independence of science and the law.

20. McComas 1998, *supra* note 4.

21. 526 U.S. 137 (1999).

22. Hull 1988, *supra* note 11; Kitcher 1995, *supra* note 5; Longino 1990, *supra* note 11; Michael Strevens, *The Role of the Priority Rule in Science*, 100 J. of Phil. 55–79 (2003), https://doi.org/10.5840/jphil2003100224; *and* Wimsatt 2007, *supra* note 5.

23. In the 1995 remand of *Daubert* (43 F.3d 1311 (9th Cir. 1995)), Judge Alex Kozinski wrote that scientific investigations performed in the service of litigation should be more stringently evaluated for admissibility, suggesting that they may not be reliable, a perspective that has been criticized by some scholars. For example, see Sheila Jasanoff, *Representation and Re-Presentation in Litigation Science*, 116 Env't Health Persps. 123–29 (2008), https://doi.org/10.1289/ehp.9976; and Leslie I. Boden & David Ozonoff, *Litigation-Generated Science: Why Should We Care?* 116 Env't Health Persps. 117–22 (2008), https://doi.org/10.1289/ehp.9987, which are largely consistent with the broad view of science as a human and community endeavor presented in this reference guide. Indeed, Jasanoff attributes

building, so long as biased methodologies and interpretations are avoided. However, research conducted for the purpose of testifying is unlikely to have had the opportunity to be reviewed and scrutinized by the relevant scientific community. Though imperfect (see section titled "Peer Review" below), this process of community scrutiny can still help to identify and overcome biases that may shape a scientific investigation, as well as assess its methodological and deductive quality (see section titled "Science as a Human and Community Endeavor" below). Because of this, judges may be placed in the position of providing a higher level of scrutiny to testimony based on research developed for litigation than that based on established research programs conducted for other purposes, which have already iterated through the flowchart multiple times. For example, survey research is an important line of evidence in trademark infringement litigation,[24] and this research may be so specific in its intent (e.g., whether the term *beanies* is perceived as a brand name for a toy[25]) that there is little reason to perform it except in service of litigation. Testimony based on such survey research is often determined to be admissible, but may not be if biases or methodological problems are identified in the study design.[26] In short, testimony based on research conducted solely for litigation may be scientifically reliable and admissible—or may not be; and making this determination in the absence of any of the services performed by the scientific community may place a higher burden on judges.

Key Traits of Science

Judges might wish for a single marker that distinguishes between science and pseudoscience, good and bad science, accepted and speculative science. Unfortunately, we have no easy answers. These are all the end points of continua, and multiple factors determine where on the spectrum a particular unit of knowledge or research falls. Contemporary thinking suggests that the boundary between science and nonscience is best identified by the *process* by which knowledge is generated, and

this denigration of litigation-derived science to resting on "idealized, misleading, or misinformed assumptions about the scientific method," which this reference guide aims to clarify and correct.

24. *See* Shari Seidman Diamond et al., *Reference Guide on Survey Research*, in this manual.

25. Ty Inc. v. Softbelly's, Inc., 353 F.3d 528 (7th Cir. 2003).

26. Rebecca Kirk Fair & Laura O'Laughlin, *Ensuring Validity and Admissibility of Consumer Surveys*, ABA Groups Practice Points (2017), https://perma.cc/2LHD-7NE2. For more on the design of methodologically sound surveys, see Shari Seidman Diamond et al., *Reference Guide on Survey Research*, in this manual.

especially by how the community engaged with that process behaves.[27] Key aspects of this process and these behaviors are outlined below.

Science Investigates the Natural World and Natural Explanations

Communities engaged in scientific endeavors aim to build accurate, natural explanations for how the world works. In this context, the term *natural* refers to any element of the physical universe—whether made by humans or not. Thus, the natural world includes matter, the forces that act on matter, energy, distant galaxies, the constituents of the biological and physical world (e.g., the ozone layer), humans, human society, and the products of that society (e.g., CFCs). As described above, the goal of building explanations does not imply that, to be scientific, research must be without practical or economic end. Both applied research, which is undertaken with the goal of solving a particular problem, and pure or basic research, which has the primary aim of building scientific knowledge, improve our understanding of the natural world.

The investigation of *super*natural explanations—that is, explanations that rely on forces beyond the observable universe—is not a part of modern science.[28] The exclusion of supernatural explanations from science was recognized in *Kitzmiller v. Dover Area School District*, in which parents of school-age children brought suit against their local school board over the requirements that Intelligent Design be taught in schools and that teachers read a disclaimer that cast evolution as shaky science.[29] The court found that Intelligent Design is not science in part because it "violates the centuries-old ground rules of science by invoking and permitting supernatural causation."[30] The philosophical and historical roots of this exclusion are deep, but a simple, practical explanation clarifies this necessity: Science is able to build reliable knowledge because it is based on evidence and observations, and supernatural explanations are generally not testable against evidence, as explained below.

27. Hull 1988, *supra* note 11; Philip Kitcher, *Believing Where We Cannot Prove*, Abusing Science: The Case Against Creationism 30–54 (1983) [hereinafter Kitcher 1983]; *and* UCMP 2022, *supra* note 12.

28. Massimo Pigliucci & Maarten Boudry, Philosophy of Pseudoscience: Reconsidering the Demarcation Problem (2013).

29. 400 F. Supp. 2d 707 (M.D. Pa. 2005).

30. The court's additional reasons for finding that Intelligent Design (ID) is not science were (1) that ID proponents falsely argue that any challenge to evolutionary theory necessarily supports ID and (2) that the scientific community has investigated and rejected ID proponents' challenges to evolutionary theory.

Science Investigates Testable Hypotheses

Communities engaged in scientific endeavors work with testable hypotheses. For a hypothesis to be testable, it must, by itself or in conjunction with other hypotheses, generate specific predictions—a set of observations that one could expect to make if the hypothesis were true and/or a set of observations that would be inconsistent with the idea and lead one to believe that it is *not* true. If an explanation is equally compatible with all possible observations, then it is not testable and hence, not within the reach of science. This is frequently the case with supernatural explanations. For example, a central proposition of Intelligent Design that a complex anatomical structure (e.g., an eye) was designed by an intelligent, supernatural entity does not make any predictions about anatomy or distribution of the structure among species.[31] We cannot know how or why a supernatural entity would design an eye in a particular way, so *any* observation we might make about the anatomical structure is compatible with this explanation. Evidence cannot be used to investigate such propositions, and hence, they are outside the realm of science.

The logic of testing hypotheses is illustrated in the central area of Figure 1. For example, the idea that smoking causes lung cancer is testable and leads to a wide variety of predictions: that rates of lung cancer will be higher among smokers than among a comparable group of nonsmokers; that lung cancer rates will increase in populations where smoking becomes common; that exposing non-human animals to compounds in cigarettes will lead to higher rates of cancer; and many more. Similarly, the hypothesis that CFCs deplete the ozone layer led to many predictions: that certain chemical reactions occur in the atmosphere; that the distribution of CFCs and other chemicals in the atmosphere will match the predictions of models; and that the ozone layer will thin as CFCs accumulate. The same logic applies to testing hypotheses much more limited in scope. For example, the narrow hypothesis that CFCs persist in the atmosphere can be tested simply by checking for them in places far from where they were produced.[32] For an example drawn from the courtroom, consider litigation surrounding whether Thermos Co.'s trademark was infringed upon by another company's use of the term *thermos*. The narrow hypothesis that *thermos* is perceived as a generic term for an insulated vacuum bottle led to the prediction that people who want to purchase an insulated vacuum bottle from a store would ask for a "thermos."[33]

In the context of testing hypotheses, the term *prediction* refers to a logical consequence of a hypothesis, not necessarily what will happen in the future. Scientific predictions often involve reexamining records of the past. For

31. Elliott Sober, *What Is Wrong With Intelligent Design?*, 82 Q. Rev. Biology 3–8 (2007), https://doi.org/10.1086/511656.

32. This hypothesis was tested and described as part of Lovelock et al. 1973, *supra* note 6, although performing that test was not the main intent of the investigation.

33. King-Seeley Thermos Co. v. Aladdin Indus., 321 F.2d 577 (2d Cir. 1963).

example, key hypotheses about connections between greenhouse gases and Earth's temperature, which have dramatic implications for life on Earth today, make predictions that are best tested, in part, by examining ice and oceanic sedimentary cores that reveal Earth's deep history.[34]

Science Responds to Evidence

Science builds knowledge based on evidence; this is how the community ensures that the explanations it generates approach accuracy. As a community, scientists actively seek evidence to test their hypotheses, and ultimately accept, give up, or modify hypotheses as warranted by the evidence. Just as Rule 702(d) directs courts to assess whether there is a disconnect between an expert's methods and principles, and the conclusions reached, reliable scientific knowledge building depends on there being a logical connection between what the evidence shows and what explanatory hypotheses are accepted. Ignoring evidence is not an option in science. For example, when research revealed the hole in the ozone layer to be much larger than any of the models predicted, scientists rushed to explore additional explanations and modifications of the key hypothesis that would help make sense of this new evidence.[35]

Evidence is not the *only* arbiter of whether a hypothesis is accepted by the scientific community. Other considerations, like broadness and coherence, also matter. Furthermore, the community's response to evidence or engagement with testing a hypothesis may be slower or hindered if dogma or other social factors work against it, as described in the section titled "Achieving Scientific Consensus" below. But evidence is the most important determinant of acceptance in science. Explanations that are not supported by evidence are eventually rejected. Hypotheses that are protected from testing or that are allowed to be tested by only one group of investigators with a vested interest in the outcome are not a part of good science.

Science Does Not *Prove* Hypotheses

Because science responds to evidence, scientific knowledge is always open to refinement, revision, or even rejection if warranted. Philosophers of science call this property *fallibilism*.[36] No matter how much evidence supports or refutes it, a

34. Valérie Masson-Delmotte et al., *Information from Paleoclimate Archives*, *in* Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change 383–464 (2013).

35. *E.g.*, Susan Solomon et al., *On the Depletion of Antarctic Ozone*, 321 Nature 755–58 (1986), https://doi.org/10.1038/321755a0.

36. Susan Haack, *Inquiry and Advocacy, Fallibilism and Finality: Culture and Inference in Science and the Law*, 2 L., Probability & Risk 205–14 (2003), https://doi.org/10.1093/lpr/2.3.205.

hypothesis cannot be absolutely proven true or false.[37] Even widely accepted scientific explanations, like the link between CFCs and ozone depletion, could be rejected if new evidence or a better-supported hypothesis comes along. This has happened at many points in the history of science. For example, Newton's classical mechanics is a powerful predictive tool that was relied upon by physicists for 200 years and is still used to generate reliable predictions in many situations. But when Einstein's theory of relativity showed itself to be more closely aligned with the evidence in more situations than Newtonian mechanics, scientists accepted that this new framework was a broader and more powerful one. At a smaller scale, the idea that one misfolded protein can cause others to misfold and generate disease (i.e., the prion hypothesis) was once regarded as beyond far-fetched and rejected by biologists; however, as evidence accrued showing prions to be proteins responsible for transmissible diseases such as mad cow disease and scrapie, dogma was overridden and scientists accepted this expanded conception of proteins and disease.[38] As new scientific facts are established and insufficient ones are overturned, judges can expect these slow-moving ripples of change to play out at trial. For example, bitemark evidence may be increasingly excluded as research establishes its fundamental limitations.[39]

Despite its tentative nature, accepted scientific knowledge is reliable. Such explanations generate predictions that hold true in many different contexts and at many different scales, allowing us to figure out how entities in the natural world are likely to behave and how we can harness that understanding to solve problems and dispense justice. For example, Molina and Rowland's understanding of chemistry allowed them to predict that CFCs were damaging the ozone layer long before this was observed. Because of this scientific knowledge, we were able to make policy changes and ultimately reverse damage to the ozone layer. We have good reason to trust accepted scientific explanations: They allow us to make accurate predictions about the future and intervene on the present. Even Newtonian mechanics, though subsumed by relativity at an explanatory level, is still used at a practical level for a wide variety of applications, such as building bridges and predicting the movement of satellites and asteroids.

37. Formal proofs are central to mathematics, a field that provides critical tools for scientists but that is nonetheless distinct from scientific disciplines that rely on evidence from the natural world to assess hypotheses. Mathematical proofs accept a set of axioms as true and demonstrate that a particular conclusion is logically implied by those assumptions. While the sciences described here incorporate logic and inference, their reliance on evidence and observations to verify conclusions means that hypotheses are fallible in a way that mathematical statements are not.

38. Claudio Soto, *Prion Hypothesis: The End of the Controversy?*, 36 Trends in Biochemical Scis. 151–58 (2011), https://doi.org/10.1016/j.tibs.2010.11.001 [hereinafter Soto 2011].

39. *See* Valena E. Beety et al., *Reference Guide on Forensic Feature Comparison Evidence*, in this manual, for more on bitemark evidence.

Science Is Carried Out by a Community that Holds Members to Norms

Communities engaged in scientific endeavors generally adhere to a set of practical behavioral standards, known as norms, which can change over time. There is no one group tasked with ensuring adherence to scientific norms, but because they are recognized as important by a wide variety of scientific institutions, scientists who are found to violate them knowingly are likely to be impeded from publishing, receiving grants, and advancing their career (see section titled "Misconduct" below for more on this). The following list summarizes some important, current scientific norms:

- *Seek relevant information:* Because science is cumulative and scientists are in the business of seeking broad explanations, scientists need to know what others before them have learned about a system to form hypotheses that cohere with the many lines of available evidence. Science rarely starts from scratch. For example, upon hearing how widespread CFCs were, Mario Molina first turned to the published work of other scientists to see what chemical processes might break down CFCs in the lower atmosphere before they reach the ozone layer.
- *Scrutinize ideas and evidence:* In science, questioning methods and looking for other explanations for evidence does not necessarily signal that a hypothesis is wrong or a study was poorly executed. For example, the prion hypothesis was once viewed suspiciously by many biologists, who required particularly convincing evidence to accept an idea that seemed at odds with how proteins were known to behave.[40] Such skepticism is part of the normal practice of science and helps steer it in the direction of accuracy. Indeed, scientists are generally motivated to scrutinize closely even well-established ideas and evidence because of the esteem that the community awards those who overturn accepted science and contribute truly new and fruitful explanations.
- *Openly communicate findings:* Scientists are expected to share their hypotheses, methods, and findings honestly and openly to allow others to evaluate and build on them. When Rowland and Molina uncovered evidence about the longevity of chlorine nitrate in the atmosphere that threatened their hypothesis, they published the evidence for others to evaluate.[41] For examples of how violation of this norm can lead to or factor into litigation, see section titled "Misconduct" below.

40. Soto 2011, *supra* note 38.

41. F. Sherwood Rowland et al., *Stratospheric Formation and Photolysis of Chlorine Nitrate*, 80 J. Physical Chemistry 2711–13 (1976) [hereinafter Rowland et al. 1976].

- *Identify and avoid bias:* Because scientists are human, it is impossible for them to be completely objective and unbiased in their work (see section titled "Science as a Human and Community Endeavor" below). However, scientific progress depends on scientists aspiring toward this ideal by *trying* to employ fair tests and methods, and to evaluate ideas on the merits of the evidence. For example, the scientists who first reported the hole in the ozone layer over Antarctica strove for objectivity in several ways. They got their data from carefully calibrated instruments, directed other scientists to information about how the machines were calibrated, and presented data with generous error bars to accommodate subsequent corrections to their observations.
- *Assimilate evidence:* When faced with evidence contradicting their hypothesis, scientists may decide to conduct more tests, may revise or reject the idea, or may consider alternative ways to explain the evidence. But ultimately, scientific explanations are sustained by evidence and cannot be propped up—or denied—if the evidence is clear. For example, as evidence mounted in favor of the prion hypothesis, even skeptics were convinced.
- *Provide appropriate credit:* Scientists are expected to cite their sources scrupulously. Credit is an important measure of impact within the scientific community. It also provides a sort of paper trail, through which another scientist can evaluate the methods, assumptions, and reasoning that a particular study is based on. Rowland and Molina's original paper on the CFC/ozone link included 31 citations outlining where their ideas and evidence originated.[42]
- *Abide by official ethical guidelines:* Research institutions, funding agencies, publishers, scientific societies, and legislation regulate what scientific research can be done and how. In general, these guidelines are intended to ensure that scientific work is of high quality, is performed in ethical ways, and benefits society. Such guidelines vary depending on the subject, application, scope, and setting of research. For example, a survey-based study designed to contribute to generalizable knowledge about human perception will generally need to be approved by an institutional review board (IRB) tasked with protecting the interests and rights of survey participants before the research can commence.

When the work of scientific experts aligns with these norms, judges can be more confident that their testimony stems from reliable scientific methodologies.

42. Molina & Rowland 1974, *supra* note 7.

Nonscience, Pseudoscience, Bad Science, and Wrong Science

The characteristics above can help us identify research that is clearly scientific, but there are different ways for an endeavor to be *un*scientific. Some lines of inquiry are simply outside the realm of science and do not pretend to be otherwise. For example, science cannot make moral judgments. Science can help us assess what segment of the population finds euthanasia acceptable, understand how people reason about euthanasia as they near the end of their lives, and learn about the emotional effects of euthanasia,[43] but science cannot tell us if euthanasia is morally right or wrong.

Pseudoscience, on the other hand, employs the trappings of science but fails to meet several key scientific standards. For example, homeopathy is a type of alternative medicine that arose in the 18th century out of legitimate concerns about the efficacy of mainstream medicine at the time. It is based on the idea that "like cures like"—that ingesting a substance that causes certain bodily effects in a healthy person (e.g., chills) can be used to cure a disease that causes that symptom. Homeopathy uses Latin scientific names for the substances in its treatments, has adopted the vocabulary of mainstream medicine (e.g., *vaccine*), and often packages and labels products in ways that are reminiscent of mainstream pharmaceuticals—all of which might be interpreted as somehow scientific. However, many evidence-based tests of homeopathic remedies have been carried out and turned up no data suggesting that they perform better than placebos. Similarly, no evidence of a natural mechanism that might underlie the claims made by homeopaths has been found. Many panels, committees, and councils tasked with assessing the scientific consensus on homeopathy have concluded that it has no scientific basis and is, instead, pseudoscience.[44] In *Allen v. Hyland's, Inc.*, the plaintiffs argued that Hyland's made false claims about the efficacy of its homeopathic remedies.[45] The defendants' expert witnesses were allowed to testify, a decision that has been criticized as a failure of the court's scientific gatekeeping function

43. For example, see Joachim Cohen et al., *European Public Acceptance of Euthanasia: Socio-Demographic and Cultural Factors Associated with the Acceptance of Euthanasia in 33 European Countries*, 63 Soc. Sci. & Med. 743 (2006), https://doi.org/10.1016/j.socscimed.2006.01.026; Ronit D. Leichtentritt & Kathryn D. Rettig, *Meanings and Attitudes Toward End-of-life Preferences in Israel*, 23 Death Stud. 323–58 (1999), https://doi.org/10.1080/074811899200993; and Nikkie B. Swarte et al., *Effects of Euthanasia on the Bereaved Family and Friends: A Cross Sectional Study*, 327 BMJ 189 (2003), https://doi.org/10.1136/bmj.327.7408.189.

44. For example, see National Health and Medical Research Council, *NHMRC Information Paper: Evidence on the Effectiveness of Homeopathy for Treating Health Conditions* (2015).

45. No. CV 12–1150 DMG (MANx) (C.D. Cal. Aug. 16, 2016); *See* Robert G. Knaier, *Homeopathy on Trial:* Allen v. Hyland's, Inc. *and a Failure of Evidentiary Gatekeeping*, 57 Jurimetrics J. 361–96 (2017).

under *Daubert*, and the jury found for the defendants, highlighting the challenges of detecting pseudoscience and assessing scientific consensus.

Then there is "bad" science: specific investigations that are within the realm of science (focus on natural explanations, work with testable ideas, etc.) but are unreliable because they have not been executed in ways that meet the community's standards for scientific behavior—perhaps because the researchers selectively reported data and so did not openly and honestly communicate findings. Some analyses suggest that cases of bad science are on the rise because of systemic incentives that reward quantity of publications and certain types of results over quality.[46] A fuller framing of this concern is provided in the section below titled "Replication and the 'Replication Crisis.'"

Unsurprisingly, bad science often leads to wrong science—that is, acceptance of hypotheses that are inaccurate. Even investigations that meet all the standards described above can wind up supporting incorrect explanations. But because science involves retesting hypotheses in different ways, contexts, and combinations, inaccurate ideas will ultimately be identified as such. This is part of the normal way in which science works, but it does take time, often many years, for scientists to sort through explanations and identify those that consistently lead to accurate predictions, new insights, and improved understanding.

Sometimes entire fields or lines of inquiry are denounced as not scientific or pseudoscientific, and such arguments may be leveraged at trial. However, as we've seen, science can address a vast array of topics. A finer-grained analysis may be required to assess *which* aspects of a body of knowledge are backed by scientific evidence and which are not. For example, acupuncture is a technique of traditional Chinese medicine used for more than 2,000 years that is based on the idea of Qi (or "Chi"), a type of life-giving energy that circulates through the body in special channels. No scientific evidence supports the existence of Qi or the channels through which it circulates. If we accept the traditional view that Qi is a mystical force, then this explanatory mechanism is *super*natural and outside the realm of science. However, sound scientific investigations have found acupuncture to be useful for treating a variety of symptoms, and scientific research is beginning to investigate the biological basis of these effects.[47] At the same time, proponents sometimes overstate these findings, and practitioners may employ terminology and tools that patients associate with medicine (often for good reason), making it difficult to distinguish which aspects of acupuncture are supported by scientific research and which are pseudoscientific. However, labeling *all* of acupuncture as nonscience ignores reliable scientific evidence and knowledge.

46. For an overview of the concern, causes, and proposed solutions, see National Academies of Sciences, Engineering, and Medicine, Reproducibility and Replicability in Science (2019) [hereinafter NASEM].

47. For example, see Tony Y. Chon & Mark C. Lee, *Acupuncture*, 88 Mayo Clinic Proc. 1141–46 (2013), https://doi.org.10.1016/j.mayocp.2013.06.009.

Science as a Human and Community Endeavor

Science is carried out not by objective automatons but by people, shaped by their unique backgrounds and motivations, as well as by the changing norms and values of the societies in which they are embedded, all of which influence the course of science. The identities, backgrounds, and experiences of the practitioners of science impact their selection of research topics, hypotheses, chosen research methods, and interpretations of results and evidence. Such influences are apparent in a wide range of scientific fields but are famously exemplified by shifts in the field of primatology as more women entered the field in the 1970s. These scientists did not make the same assumptions that their male predecessors had made and so observed and investigated behaviors and interactions previously overlooked and undocumented, revealing, for example, that female primates have elaborate sex lives and manipulate male behavior.[48] Of course, the fact that scientists are fallible humans also means that they are vulnerable to bias and capable of misconduct, as described in the sections below.

The human practitioners and institutions of science likewise respond to events in and values of society around them in various ways. For example, the 1970s energy crisis led to a surge of alternative energy research and reorganizations of funding mechanisms for this research.[49] To cite an even broader

48. For a few examples of how scientists' identities, backgrounds, and experiences impact their science, across several disciplines, including in the field of primatology, see Donna Haraway, *Primate Visions: Gender, Race, and Nature in the World of Modern Science* (1989); Andrew Gary Darwin Holmes, *Research Positionality—A Consideration of Its Influence and Place in Qualitative Research—A New Researcher Guide*, 8 Shanlax Int'l J. of Educ. 1–10 (2020), https://doi.org/10.34293/education.v8i4.3232; Rembrand Koning, Sampsa Samila, & John-Paul Ferguson, *Who Do We Invent For? Patents by Women Focus More on Women's Health, but Few Women Get to Invent*, 372 Science 1345–48 (2021), https://doi.org/10.1126/science.aba6990; Diego Kozlowski et al., *Intersectional Inequalities in Science*, 119 Proc. of the Nat'l Acad. of Sci. USA e2113067119 (2022), https://doi.org/10.1073/pnas.2113067119 [hereinafter Kozlowski et al., 2022]; D. N. Lee, *Taking it Personally,* 311 Sci. Am. 47 (2014), https://doi.org/10.1038/scientificamerican1014-47; Douglas Medin, Carol D. Lee, & Megan Bang, *Particular Points of View*, 311 Sci. Am. 44–45 (2014), https://doi.org/10.1038/scientificamerican1014-44; H. Richard Milner, IV, *Race, Culture, and Researcher Positionality: Working through Dangers Seen, Unseen, and Unforeseen*, 36 Educ. Researcher 388–400 (2007), https://doi.org/10.3102/0013189X07309471; Mathias Wullum Nielsen et al., *One and a Half Million Medical Papers Reveal a Link Between Author Gender and Attention to Gender and Sex Analysis*, 1 Nature Hum. Behav. 791–96 (2017), https://doi.org/10.1038/s41562-017-0235-x; Cassidy R. Sugimoto et al., *Factors Affecting Sex-Related Reporting in Medical Research: A Cross-Disciplinary Bibliometric Analysis*, 393 Lancet 550–59 (2019), https://doi.org/10.1016/S0140-6736(18)32995-7; Caitlin Wilson, Gillian Janes, & Julia Williams, *Identity, Positionality and Reflexivity: Relevance and Application to Research Paramedics*, 7 Brit. Paramedic J. 43–49 (2022), https://doi.org/10.29045/14784726.2022.09.7.2.43.

49. Alice L. Buck, *History of the Energy Research and Development Administration*, No. DOE/ES-0001 USDOE Assistant Secretary for Management and Administration (1982).

example, while the earliest examples of scientific peer review date back hundreds of years, scholarship has traced the rise of peer review as a central pillar of scientific publication and funding—as well as its perception as a critical hallmark of "good" science—to the Cold War. That conflict instigated an expansion of public monetary support for scientific research and, as society's anxieties about American scientific competitiveness began to ease, led to a tension between public accountability and scientists' personal desires for autonomy, respect, and financial support.[50] Individual scientists and the overarching institutions of science are shaped by society and, in turn, produce scientific knowledge that goes on to affect society through its applications and implications, creating the feedback loop between "benefits and outcomes" and other parts of the process of science illustrated in Figure 1. Both scientific institutions and scientists are deeply embedded within and interconnected with broader society.

Social interactions within the scientific community are critical to building scientific knowledge. We saw this in the CFC/ozone example as one researcher's conference presentation sparked Molina and Rowland's first investigation of CFCs, which, in turn, led to a cascade of studies by different research groups. Scientists often work collaboratively, and today more and more science is done by large research groups and multi-institution consortia, reflecting the increasing complexity of scientific questions being investigated and the specialized knowledge needed to carry out those investigations. The scientific community performs other key services as well. One of the most important of these is scrutinizing the evidence and ideas of other scientists.

Scientific scrutiny often involves the following activities, which help ensure the reliability of the scientific knowledge produced:

- an evaluation of methods, considering potential biases and oversights,
- reconsideration of the hypothesis and how to interpret the evidence relevant to it, and
- appraisal of alternative hypotheses that may also explain the results.

Scientists engage in these activities at many times, including when serving as peer reviewers for a publication, when reviewing research for their own information, and through questions and discussions at conferences. In their published articles, scientists generally preserve an account of how they have applied this scrutiny to others' work and to their own research described in the article. For example, a key paper that helped show that proteins are the infectious agents in prion diseases documents the careful and often laborious reasoning required

50. Melinda Baldwin, *Scientific Autonomy, Public Accountability, and the Rise of "Peer Review" in the Cold War United States*, 109 Isis 538–58 (2018) [hereinafter Baldwin 2018].

to fully scrutinize a hypothesis and the relevant evidence.[51] In that paper, the author critiques other researchers' methods of purifying the infectious agent and explains why they might not be useful; considers fourteen different previously proposed alternative hypotheses regarding the nature of the infectious agent in prion diseases; outlines the many different lines of evidence relating to these hypotheses; considers modifications of these hypotheses that might still account for the available data; describes where the author's findings differ from those of other investigators; explains how one set of data the author collected is consistent with multiple hypotheses; acknowledges when clearly relevant factors could not be taken into account in the author's calculations; and describes where more research needs to be done because of known problems with certain assays. The application of such a high degree of scrutiny reflects the goal of producing accurate explanations for the natural world despite the fundamentally fallible nature of scientific knowledge. When research relevant to a trial has not yet been scrutinized by a community with the appropriate technical expertise, a judge may be placed in the position of providing or requesting this scrutiny.

Ultimately, through such scrutiny, scientists decide what ideas, results, and methods are worth building on and retesting in the same or some other context. Because scientists are unique, societally situated, and fallible humans, the scientific community can better perform the functions listed above, as well as many others, when scientists represent the diversity of the societies in which science is embedded.[52] A diverse scientific community balances biases, investigates areas of inquiry relevant to all parts of society, and explores more innovative ideas for addressing the challenges it sets for itself.[53] Science is making progress toward equitable inclusion, albeit slowly.[54]

51. Stanley B. Prusiner, *Novel Proteinaceous Infectious Particles Cause Scrapie*, 216 Science 136–44 (1982), https://doi.org/10.1126/science.6801762.

52. Here, the term *diversity* includes racial and gender identity, of course, but also many other facets of identity and background, including culture, religion, age, sexual orientation, disability, incarceration history, class, and more.

53. For evidence that diversity fosters the investigation of diverse questions of relevance to society, see Travis A. Hoppe et al., *Topic Choice Contributes to the Lower Rate of NIH Awards to African-American/Black Scientists,* 5 Sci. Advances 1–12 (2019), https://doi.org/10.2226/sciadv.aaw7238, and Kozlowski et al., 2022, *supra* note 48. For evidence that diversity fosters innovation, see Bas Hofstra et al., *The Diversity-Innovation Paradox in Science*, 117 PNAS 9284–91 (2020), https://doi.org/10.1073/pnas.1915378117. For more on the mechanisms behind this connection, see Patrick Grim et al., *Diversity, Ability, and Expertise in Epistemic Communities*, 86 Phil. Sci. 98–123 (2019), https://doi.org/10.1086/701070.

54. For more on changes in the diversity of scientific workforce in the U.S., see Pew Rsch. Ctr. *STEM Jobs See Uneven Progress in Increasing Gender, Racial, and Ethnic Diversity* (2021). For a global picture, see Fred Guterl, *Where Are the Data?*, 311 Sci. Am. 40–41 (2014).

Peer Review

Peer-reviewed journal articles are the most important way that scientists convey their ideas and evidence to the scientific community, offering them up for scrutiny. Indeed, in *Kitzmiller v. Dover Area School District*, the court found that Intelligent Design could not be considered an accepted scientific theory in part because its proponents had not published research in peer-reviewed journals. That is not to suggest that all testimony at trial must be based on peer-reviewed research, but that information required to be addressed in the science classroom should meet this minimum standard. *Daubert*, of course, includes peer review as one of five nonexclusive factors judges may consider in assessing the scientific validity of expert testimony.

In the process of peer review, editors at a journal, typically themselves scientists, receive a submission and first decide if it is appropriate for publication in that journal: whether it is sound enough, on topic, and of sufficient importance. If they determine that it is, they send it to other scientists with expertise on the topic of the article—the peer reviewers—who are not known to have a conflict of interest: for example, scientists who are not direct collaborators of the paper's authors or identified by the authors as potentially vested in the publication decision. The peer reviewers may recommend rejection or acceptance of the paper outright or may request changes, starting a back-and-forth among the authors, reviewers, and editors until the editors make a final decision on acceptance or rejection. Peer review and publication are time-consuming, usually involving many months between submission and publication. The process can also be highly competitive. The journal *Science* accepted just 6.4% of the articles it received in 2021.

When the institution of peer review was adopted by the scientific community, it was not intended to distinguish credible science from that which is inaccurate or badly executed, though the process is now often viewed as serving that function.[55] Nevertheless, peer review is an important gatekeeper to the official record of science, ensuring that published works at least purport to meet minimum standards of quality and objectivity. In this way, it is reasonable to consider whether scientific knowledge that might impact a trial has been published in a peer-reviewed journal. However, peer-reviewed publication is a far from surefire indicator of reliable science. Here are some of the reasons.

- Peer-reviewed journals vary in their standards and criteria for publication. Some are much more lax about methodological rigor than others, and often these differences are readily apparent only to members of the expert scientific communities themselves. Judges may be tempted to rely on a journal's impact factor, a statistic that indicates how often articles

55. Baldwin 2018, *supra* note 50.

published in that journal are cited and hence loosely conveys its prestige within a field, as an indication of methodological rigor. However, prestige is not a simple corollary of quality or correctness. A journal may have high standards for publication but a lower impact factor because of its topic or the type of articles it publishes (e.g., review articles are likely to garner more citations). Furthermore, journals can use various strategies to artificially boost their impact factor.[56] To understand the quality of a particular specialized journal, a judge may need to rely on experts within the field.

- Most but not all of the material that is published in peer-reviewed journals is original research. In addition to research articles, such journals publish commentary, including editorials, that may or may not be peer reviewed and that may express opinions and perspectives not immediately derived from evidence, and identifying these commentaries as such can sometimes present a challenge to nonspecialists.
- Recognizing the weight that peer-reviewed publications are given, sometimes regardless of the quality of the evidence they present, proponents of pseudoscientific or rejected lines of inquiry have begun to put out peer-reviewed journals that are oriented toward promoting a particular viewpoint, not objectively weighing different lines of evidence.
- Peer reviewers usually cannot detect scientific fraud. For example, widely cited research investigating the cause of Alzheimer's disease published in the prestigious peer-reviewed journal *Nature* in 2006 allegedly includes fraudulent, manipulated images.[57] When scientific fraud is discovered, it may result in retraction (see section titled "Correction and Retraction" below), but this process of discovery and resolution takes time. The validity of the 2006 Alzheimer's disease paper's results was not brought into question until 2021, when an unaffiliated scientist was commissioned to investigate the research by an attorney for investors.[58] Then in 2022, *Nature*'s editors attached a note to the publication cautioning readers that concerns had been raised and that an investigation was ongoing. The paper was retracted on June 24, 2024.
- Many investigations that appear in peer-reviewed journals may reach an incorrect conclusion, which is revealed only when further testing is applied to the hypothesis. This is a normal part of the process of science.

56. Douglas N. Arnold & Kristine K. Fowler, *Nefarious Numbers*, 58 Notices Am. Mathematical Soc'y 434–37 (2011), https://doi.org/10.48550/arXiv.1010.0278.

57. Sylvain Lesné et al., *A Specific Amyloid-β Protein Assembly in the Brain Impairs Memory*, 440 Nature 352–57 (2006), https://doi.org/10.1038/nature04533 [hereinafter Lesné et al. 2006].

58. Charles Piller, *Blots on a Field?*, 377 Science 358–63 (2022), https://doi.org/10.1126/science.add9993.

- Despite efforts to eliminate it, bias may influence the peer review of individual papers, as well as editors' decisions about publishing those papers. Some have argued that the process may be conservative and favor established hypotheses and less innovative approaches.[59] Others have suggested that the process currently favors certain types of positive results, even if they turn out to be incorrect.[60] And studies have reported different findings in terms of whether historically oppressed groups experience bias in the peer-review process.[61]

Rowland and Molina's CFC hypothesis was first published in *Nature*, which aims to publish original work that is of widespread interest and importance across science. Rowland and Molina's work met both of these requirements and ultimately led to a widely accepted scientific hypothesis. One might think that if a journal like *Nature* publishes a paper, its hypotheses are more likely correct, at least in comparison with hypotheses published in less-competitive journals. However, because such journals require that the work they publish be of widespread interest and importance, in fact, the papers that appear in them may be more likely to concern "risky" hypotheses that make bold explanations challenging established ideas or that leverage new techniques still undergoing development. In addition, "everyday" science, which applies widely accepted methods in routine ways, is unlikely to warrant publication in preeminent journals. Journal prestige is thus not a clear indicator of scientific reliability. We saw that Rowland and Molina's hypothesis was basically correct when it was published in *Nature*, but it still underwent many refinements and modifications afterward, particularly to the mechanisms involved in CFCs' depletion of the ozone layer.

Some new research may be on track to undergo peer review but be so recent that this has not yet taken place. Such research is often shared in the form of preprints, fully drafted articles that include all the features of scientific journal articles but have not yet been accepted to a peer-reviewed journal for publication.[62] Preprints are often shared via online repositories (often with names of the format arXiv, bioRxiv, medRxiv, etc.). Such repositories may accept or

59. Kyle Siler, Kirby Lee, & Lisa Bero, *Measuring the Effectiveness of Scientific Gatekeeping*, 112 PNAS 360–65 (2014), https://doi.org/10.1073/pnas.1418218112.

60. For example, see Paul E. Smaldino & Richard McElreath, *The Natural Selection of Bad Science*, 3 Royal Soc'y Open Sci. 160384 (2016), https://dx.doi.org/10.1098/rsos.160384.

61. For a few examples, see Donna K. Ginther et al., *Race, Ethnicity, and NIH Research Awards*, 333 Science 1015–19 (2012), https://doi.org/10.1126/science.119783; Markus Helmer et al., *Gender Bias in Scholarly Peer Review*, 6 eLife e21718 (2017), https://doi.org/10.7443/eLife.21718; Patrick S. Forscher et al., *Little Race or Gender Bias in an Experiment of Initial Review of NIH R01 Grant Proposals*, 3 Nature Hum. Behav. 257–64 (2019), https://doi.org/10.1038/s41562-018-0517-y; and Flaminio Squazzoni et al., *Peer Review and Gender Bias: A Study on 145 Scholarly Journals*, 7 Sci. Advances 1–11 (2021), https://doi.org/10.1126/sciadv.abd0299.

62. *See* Oya Y. Rieger, Preprints in the Spotlight: Establishing Best Practices, Building Trust (2020), https://doi.org/10.18665/sr.313288.

reject submissions, but the process behind these determinations may not be disclosed and is not standardized in the way that peer review typically is. Most investigators who submit to preprint repositories *intend* for the work to undergo peer review and publication eventually but may not be successful at that. The fact that an article is called a *preprint* does not mean that publication is imminent. Preprints are beginning to make their way into the courts. For example, in a recent case, an incarcerated person moved to be granted compassionate release based on the risks he faced if he were to contract Covid-19.[63] The court denied the motion in part because the plaintiff had been vaccinated, mitigating his risk, and cited several lines of evidence supporting this view, including a preprint examining the efficacy of the Moderna vaccine against different strains of SARS-CoV-2.[64]

Furthermore, some technical research relevant to trials, particularly that undertaken for the purpose of litigation and that conducted within industry to support functions of narrow interest, will not have had the opportunity to undergo peer review because of its provenance and intent. As addressed in *The Admissibility of Expert Testimony* in this manual, the fact that research has not undergone peer review does not make testimony based on it inadmissible; however, it may place a higher burden on the experts to explain their methods—and on the judge to assess the validity of this explanation.

Bias

As described above, scientists hold each other to a set of norms that help science build accurate explanations. However, scientists are human beings and, while they strive toward objectivity, they also bring their unique backgrounds and experiences to bear on their work. This invigorates problem solving, sparks new ideas, and benefits science in many ways. But it also means that the unconscious and conscious biases we all hold can influence the course of science and that scientists may interpret the same data in different ways.[65] The norms of scientific behavior, for example the expectations of scrutiny and open communication, as well as processes involving the scientific community like the institution of peer review, serve the function of helping identify and correct biases. Scientific objectivity thus lies in a diverse community of inquirers, not individual scientists.

63. United States v. Singh, 525 F. Supp. 3d 543 (M.D. Pa. 2021).

64. Kai Wu et al., *mRNA-1273 Vaccine Induces Neutralizing Antibodies Against Spike Mutants from Global SARS-CoV-2 Variants,* bioRxiv https://perma.cc/SU3B-9ZCZ.

65. Heather E. Douglas, Science, Policy, and the Value-Free Ideal (2009), https://doi.org/10.2307/j.ctt6wrc78; Hugh Lacey, Is Science Value Free? (2004), https://doi.org/10.4324/9780203983195; *and* Miriam Solomon, Social Empiricism (2007).

In examining the methodologies and conclusions of expert testimony, the source of funding for the research is a valid consideration, though not in and of itself a reason to exclude testimony. Much evidence suggests that funding source is correlated with study outcome.[66] For example, in pharmaceuticals, there is a strong tendency for industry-sponsored trials to favor the industry's product.[67] Several possible, non–mutually exclusive explanations may underlie this general effect, and they are not all nefarious—e.g., perhaps drug companies mainly fund clinical trials when internal research already strongly supports a compound's efficacy. However, biased study design or biased interpretation directly introduced by the funder or by a potentially unconscious sense of quid pro quo on the part of the researcher is also a possibility.

Many potential sources of bias can be subtle and/or hard to detect. Within this manual, the reference guides on forensic science, statistics and research methods, survey research, and epidemiology outline many ways that methods and analytic techniques can subtly bias research outcomes and, hence, testimony. And of course, research sponsored by a defendant or plaintiff that fails to support the sponsor's interest or that turns out to support the interests of the other party may simply not be introduced at trial.[68] In other cases, the influence of sponsorship is obvious. For example, after Molina and Rowland's hypothesis began to gain traction, a CFC-industry-backed group sponsored a speaking tour by a meteorologist who considered the hypothesis "utter nonsense"—an approach that seemed to be recognized by the media for what it was, a public relations stunt.[69]

That is not to suggest that government- or nongovernmental organization (NGO)-sponsored research is necessarily free from bias. Individuals, academic

66. For reviews, see Sheldon Krimsky, *Do Financial Conflicts of Interest Bias Research? An Inquiry into the "Funding Effect" Hypothesis*, 38 Sci., Tech., & Hum. Values, 566–87 (2012), https://doi.org/10.1177/0162243912456271; and David B. Resnik & Kevin C. Elliott, *Taking Financial Relationships into Account When Assessing Research*, 20 Accountability Rsch. Policies & Quality Assurance, 184–205 (2013), https://doi.org/10.1080/08989621.2013.788383.

67. Sergio Sismondo, *Pharmaceutical Company Funding and Its Consequences: A Qualitative Systematic Review*, 29 Contemp. Clinical Trials 109–13 (2008), https://doi.org/10.1016/j.cct.2007.08.001.

68. On the other hand, failing to conduct or report on a study that a litigant could have conducted may be viewed with suspicion, as noted in *Eagle Snacks, Inc. v. Nabisco Brands, Inc.*, where the court observed that "failure of a trademark owner to run a survey to support its claims of brand significance and/or likelihood of confusion, where it has the financial means of doing so, may give rise to the inference that the contents of the survey would be unfavorable. . . ." 625 F. Supp. 571 (D.N.J. 1985).

69. The Council on Atmospheric Sciences sponsored the tour by British scientist Richard Scorer. Lydia Dotto & Harold Schiff, The Ozone War (1978) [hereinafter Dotto & Schiff 1978]; Edward A. Parson, Protecting the Ozone Layer: Science and Strategy (2003) [hereinafter Parson 2003]; *and* Walter Sullivan, *Scientist Doubts Spray Cans Imperil Ozone Layer*, N.Y. Times (July 8, 1975).

institutions, and other organizations may also have conflicts of interest (COIs) that are not immediately apparent—for example, through stock ownership. Government-sponsored researchers may feel motivated to secure their next grant with a result that will impress or that aligns with an agency's interests or policy positions. In an example that goes beyond mere bias, the potentially fraudulent Alzheimer's research described above was funded by a government agency, the National Institutes of Health, and was carried out by researchers mainly associated with universities.[70] Nevertheless, at the time of publication of that paper, two of the authors were listed as inventors on a patent application stemming from the research submitted by the public university that employs them. Clearly, receiving sponsorship from a government agency and being situated in an academic setting is no guarantee that research is untainted. All research is potentially influenced by bias, and *every* funder of research has the potential to introduce a source of bias. Evidence has highlighted the strength of this connection when it comes to industry-sponsored research. For this reason, disclosure of funding sources and COIs is generally a requirement of peer-reviewed publication. Molina and Rowland, for example, noted in their original paper that it was funded by the U.S. Atomic Energy Commission.[71] Some scientific institutions have COI policies designed to discourage research that they view as more likely to be biased, and many have COI-disclosure policies designed to encourage scientists to take sponsorship into account when evaluating each other's work. An examination of an investigation's funding source may help judges be alert to potential biases and assess their likely direction.

Misconduct

Beyond bias, scientific misconduct can occur when a scientist doesn't fairly evaluate other scientists' work, doesn't honestly report methods and results, doesn't fairly assign credit, or doesn't work within the ethical guidelines of the community.[72] These deceitful practices are penalized and discouraged in various ways, including with career curtailment, funding bans, fines, and potentially prison time. Scientific misconduct, as well as honest mistakes, can lead to the acceptance of inaccurate hypotheses. Such errors will be corrected as ideas are reexamined in different contexts, including in the course of a trial. For example, in *Schueneman v. Arena Pharms., Inc.*,[73] the court found that a pharmaceutical company had engaged in misrepresentation when it claimed that animal studies showed its

70. Lesné et al., 2006, *supra* note 57.

71. Molina & Rowland 1974, *supra* note 7.

72. For more, see Charles Gross, *Scientific Misconduct*, 67 Ann. Rev. of Psych. 693–711 (2016), https://doi.org/10.1146/annurev-psych-122414-033437.

73. 840 F.3d 698 (9th Cir. 2016).

drug to be noncarcinogenic, when in fact the company was conducting a study that linked the drug to cancer in lab rats, a case in which scientific results were not honestly reported. Another well-known example of scientific misconduct on trial involves Takata Corporation submitting fraudulent data on the performance of its airbag inflators to automakers. Takata pled guilty to wire fraud in 2017, nearly 20 years after its internal research had shown that its inflators did not perform to required specifications.[74]

Scientific misconduct can be difficult to detect; there are no simple red flags that a judge could identify that would indicate misconduct.[75] For example, misconduct may be reported by someone with firsthand knowledge of the environment in the research group, it may be detected by applying specialized tools for image analysis to publications, or it may be detected by subject matter experts who notice a discrepancy or anomaly in reported data or cannot replicate a result. An allegation of misconduct will generally lead to an extensive investigation and may take many years to resolve and correct. A judge or jury might justifiably be concerned that a scientist who has engaged in documented misconduct may be an unreliable source of expert testimony.

Correction and Retraction

After peer review and publication, a scientific paper may be corrected if a small error that does not invalidate the main findings of the paper is caught—for example, an incorrect row in a data table, missing text, or a missing citation. In such cases, a publisher will generally update the version of record available online and associate a correction notice with the document so that readers have a record of what changes have been made and why. Guidelines and policies for error corrections may vary across journals and are generally intended to maintain the validity of the scientific record, though it has been argued that journals and editors may not have made sufficient efforts in this regard.[76] The fact that a paper has had a correction made should not be taken to indicate that the study is of poor quality. Instead, it suggests that the article has received additional scrutiny after publication and that the authors and publishers have made public efforts to address concerns

74. *In re* Takata Airbag Prods. Liab. Litig., 193 F. Supp. 3d 1324 (S.D. Fla. 2016).

75. The Office of Research Integrity of the U.S. Department of Health and Human Services has identified five red flags of research misconduct, but all are observations that would only be made by other scientists working in the same field or by other workers with a professional relationship to the scientist suspected of misconduct. Office of Research Integrity, *Possible Red Flags of Research Misconduct* (2016), https://perma.cc/JA9J-XKTE.

76. Lonni Besançon et al., *Correction of Scientific Literature: Too Little, Too Late!*, 20 PLoS Biology e3001572 (2022), https://doi.org/10.1371/journal.pbio.3001572; Ambar Castillo, *Mistakes Happen in Research Papers. But Corrections Often Don't*, STAT (Jan. 10, 2023), https://perma.cc/TQX7-3YXW.

raised by that scrutiny, which can be viewed as a strength. Corrections in scientific papers are rare and have remained relatively constant since the 1970s.[77]

On the other hand, if evidence arises that the entire basis of a published study is invalid, the paper may be retracted. Articles may be retracted because of a significant flaw that changes the conclusions of the paper (e.g., faulty equipment used to collect data or an error in code used to analyze the data) or misconduct of many varieties—plagiarism, fraudulent data, ethics violations, or manipulations of the peer-review process.[78] Theoretically, retraction means removal from the scientific record. In practice, however, the article is likely to remain available in libraries and online tagged with a notice of retraction, meant to signal to readers that the journal no longer backs the content of that article. One reason these papers remain available is because of the cumulative nature of science: Researchers may need to review retracted papers to see if the errors within them have propagated into other work. Because such articles remain accessible, retracted research may continue to garner citations improperly.[79] While the number and frequency of retractions from scientific journals has been on the rise, multiple analyses suggest that this is largely due to improved oversight and new tools to detect plagiarism and image manipulation.[80] Studies are retracted because they do not emerge from sound scientific methodologies, and thus any science that has been retracted should be disallowed in expert witness testimony.

Within science, retractions may be viewed as a blot on one's record of scholarship, and the scientist behind a retracted study may experience stigma from the scientific community.[81] Should the same be true in the courtroom? Retraction is not the same as scientific misconduct, in which there is an intent to deceive. While scientific misconduct is the most frequent cause of retraction, at least in the life sciences, it is by no means the only cause.[82] The circumstances that can lead to

77. Daniele Fanelli, *Why Growing Retractions Are (Mostly) a Good Sign*, 10 PLoS Med. e1001563 (2013), https://doi.org/10.1371/journal.pmed.1001563 [hereinafter Fanelli 2013].

78. William Bülow, et al., *Why Unethical Papers Should be Retracted*, 47 J. Med. Ethics (2021), https://doi.org/10.1136/medethics-2020-106140.

79. Tzu-Kun Hsiao & Jodi Schneider, *Continued Use of Retracted Papers: Temporal Trends in Citations and (Lack of) Awareness of Retractions Shown in Citation Contexts in Biomedicine*, 2 Quantitative Sci. Stud. 1144–69 (2022), https://doi.org/10.1162/qss_a_00155.

80. Jeffrey Brainard & Jia You, *What a Massive Database of Retracted Papers Reveals About Science Publishing's 'Death Penalty'*, Science (Oct. 25, 2018), https://doi.org/10.1126/science.aav8384; Fanelli 2013, *supra* note 77; *and* Ferric C. Fang, R. Grant Steen & Arturo Casadevall, *Misconduct Accounts for the Majority of Retracted Scientific Publications,* PNAS Early Edition (2012) [hereinafter Fang et al. 2012].

81. This stigma may be detrimental to scientific knowledge building because it discourages the reporting of honest errors; for example, see Jaime A. Teixeira da Silva & Aceil Al-Khatib, *Ending the Retraction Stigma: Encouraging the Reporting of Errors in the Biomedical Record*, 17 Rsch. Ethics 251–59 (2021), https://doi.org/10.1177/1747016118802970.

82. Fang et al. 2012, *supra* note 80.

retraction include honest mistakes (which may initially be known only to the researcher who made them), reported out of respect for the knowledge-building function of science. In such cases, acknowledgment of the error might reasonably be perceived as a sign of integrity and commitment to developing accurate scientific explanations. For example, in 2006, pharmaceutical scientist Geoffrey Chang and colleagues retracted five papers that relied upon a flawed computer program, a mistake they owned up to and corrected.[83] Chang then continued his research, making valuable contributions to our understanding of microbial resistance, malaria, and crop development, with no indications that his new findings should be discounted because of a previous error. However, a pattern of retracted research stemming from a single researcher or group and tied to allegations of misconduct or bad science *can* suggest that those practitioners are not adhering to the norms of the scientific community, making their unretracted research suspect as well.

Retraction does not occur if the hypothesis a study supports is simply found to be incorrect by subsequent research. The methods and data of that study may still hold valuable insights for researchers, who can go back to it to try to understand why a different conclusion was reached in that case. This is a normal part of the self-correcting process of science.

Scientific Expertise

Judges may need to evaluate the qualifications and expertise of individual scientists to help determine whether to admit scientific testimony. On what basis can this judgment be made? Scientific experts are respected and active participants in their scientific communities, usually contributing to knowledge development as researchers and certainly scrutinizing the work of others as consumers of the research literature. Just as there is no single trait that separates science from nonscience, there is no single necessary and sufficient condition that would qualify a scientist as an expert on a particular topic in the eyes of other scientists. Instead, judges may evaluate a cluster of traits that a scientist is likely to attain in the process of contributing to a field and participating in the scientific community. These include:

- holding an advanced degree, usually a Ph.D., in a field closely related to the area of purported expertise,
- holding a position in which reviewing the latest research on the topic is required (e.g., having a specialized medical practice, an editorial position at a scientific journal, or a research career),

83. Greg Miller, *A Scientist's Nightmare: Software Problem Leads to Five Retractions*, 314 Science 1856–57 (2006), https://doi.org/10.1126/science.314.5807.1856.

- holding leadership positions in relevant scientific societies or institutions, or having other markers of professional esteem by the community,
- having published peer-reviewed research on the topic, and
- developing a track record, often measured in citation metrics, of publications that are well respected by colleagues and contribute to knowledge development.

Citation metrics, or citation impacts, are statistics that indicate how often a particular academic work, author, or journal has been cited. These metrics range from simple tallies to more complex statistics, such as the h-index, which incorporates information about the distribution of citations across an author's publication record.

Lacking a few of the traits in the list above does not disqualify an individual as a scientific expert, but lacking all or most of them should lead one to question whether the proposed expert is indeed an active, expert member of the relevant scientific community. Sherwood Rowland and Mario Molina had accumulated all of these markers of expertise by the end of their careers, of course, but Molina was an early career scientist at the time the CFC/ozone hypothesis was published, with only a few publications and a Ph.D. Had he been proposed as an expert witness on the ozone layer at that time, a judge might reasonably have questioned his expertise.

It is also worth evaluating the closeness of a scientist's disciplinary expertise to the scientific topic on which expert testimony will be delivered. Doctors and honorifics abound, and promoters of pseudoscientific ideas or hypotheses with little evidence supporting them may leverage achievements in other areas to create the impression of relevant expertise. For example, Richard Scorer, a well-known scientist, was promoted as an expert on the CFC/ozone hypothesis by an industry-backed group aiming to discredit the hypothesis.[84] Scorer had a Ph.D., had received scientific honors, and had conducted research on air pollution; however, he had no expertise in atmospheric chemistry or the ozone layer and had published no peer-reviewed articles on this topic. Scorer was, in fact, an expert in some related disciplines, but lacked the specialized knowledge needed to fairly evaluate the CFC/ozone hypothesis in light of the available evidence and research. The problem of scientists with legitimate expertise in one field weighing in on a scientific question outside their area of expertise is a pernicious one that has affected public acceptance of science and policy on issues such as climate change and tobacco exposure.[85] This scenario may play out in courts as judges identify and disallow testimony from scientists whose expertise is legitimate but not relevant to the specific scientific issue at hand. Many of the later reference

84. Dotto & Schiff 1978, *supra* note 69; Parson 2003, *supra* note 69.
85. Oreskes & Conway 2010, *supra* note 2.

guides in this manual provide guidance on qualifications for experts in particular disciplines.

Testing Hypotheses

At the most basic level, scientists test their ideas by figuring out what predictions are generated by a hypothesis and comparing those to evidence. Hypotheses are supported when the evidence matches predictions and are contradicted when they do not match.[86] Observations are often made with the help of sophisticated tools and may be analyzed with statistical techniques to discern patterns. Other observations are entirely qualitative and nonnumeric, taking the form of case studies or interviews, for example. Here, we review some overarching approaches to testing hypotheses that cut across all disciplines and address common concerns relating to the interpretation of data from such tests.

Scientific Methodologies and Data

The methodologies used in different scientific disciplines are so varied that it might seem a challenge to identify any fundamental similarities. How can it be that investigations as diverse as measuring ozone levels over Antarctica, exposing mice to cigarette smoke tars, and surveying consumers about their perception of the word *thermos* all serve the same fundamental purpose of testing hypotheses? In fact, scientific methods used to test hypotheses can be grouped in a few common ways, and understanding these can help in evaluating the strengths of individual studies. These attributes include:

- whether or not the research involves an intentional manipulation of condition (i.e., Is the research experimental or nonexperimental in nature?)
- what types of data the investigation produces (i.e., Did the research generate quantitative data, qualitative data, or both?)
- whether or not the hypotheses being examined are reflected in a mathematical model.

There are no right, wrong, better, or worse answers to the above questions. Instead, these attributes help determine the types of hypotheses and approaches best suited to gather evidence, as described below. Note that these characteristics vary relatively independently of each other and of scientific discipline: Experimental or nonexperimental methods can be used to collect quantitative or

86. Michael Strevens, The Knowledge Machine: How Irrationality Created Modern Science (2020).

qualitative data, with or without incorporating modeling, and these methodologies are applied broadly across the many topics that science investigates. Later reference guides in this manual explain how these methodological approaches are used in various disciplines.

Experimental and Nonexperimental Methods

An experiment involves intentionally manipulating some factor in a system to learn how that affects the outcome.[87] A controlled experiment seeks to keep variables other than the experimental manipulation the same in order to isolate the cause of any change. A control group, on the other hand, is a comparison condition or group that does not receive the experimental manipulation to increase confidence that the change observed in the experimental group is caused by the manipulation and not some other factor. For example, experiments examining whether the drug Bendectin causes birth defects, the basis of the *Daubert* litigation, compared experimental groups of animals given Bendectin during pregnancy to control groups of animals not given the drug.[88] Experiments can be performed in a lab or in a natural setting. For example, economic decision making may be investigated by bringing participants into a psychology lab and observing them playing a game, or by doing a field experiment at a market or swap meet.[89]

A randomized controlled trial is an experimental approach that randomly assigns participants to the experimental or control group to better control variables across the groups. It is most commonly used in medical trials, but can be applied more broadly as well, for example to examine the effectiveness of a kindergarten curriculum at reducing bullying.[90] Randomized controlled trials can be very convincing and are often perceived as the gold standard in medicine.[91] However, they are just one line of evidence among many.

87. Nancy Cartwright, Nature's Capacities and Their Measurement (1989).

88. For an example, see Rochelle W. Tyl et al., *Developmental Toxicity Evaluation of Bendectin in CD Rats*, 37 Teratology 539 (1988), https://doi.org/10.1002/tera.1420370603. For an overview, see Joseph Sanders, *From Science to Evidence: The Testimony on Causation in the Bendectin Cases*, 46 Stanford L. Rev. 1–86 (1993), https://doi.org/10.2307/1229235.

89. For example, see Werner Güth & Oliver Kirchkamp, *Will You Accept Without Knowing What? The Yes-No Game in the Newspaper and in the Lab*, 15 Experimental Econ. 656–66 (2012), https://doi.org/10.1007/s10683-012-9319-7; and Steven D. Levitt & John A. List, *Field Experiments in Economics: The Past, the Present, and the Future*, 53 Eur. Econ. Rev. 1–18 (2009), https://doi.org/10.1016/j.euroecorev.2008.12.001.

90. For example, see Adele Diamond et al., *Randomized Control Trial of* Tools of the Mind*: Marked Benefits to Kindergarten Children and Their Teachers*, 14 PLoS ONE e0222447 (2019), https://doi.org/10.1371/journal.pone.0222447.

91. For example, see *In re Bextra & Celebrex Mktg. Sales Practices & Prod. Liab. Litig.,* 524 F. Supp. 2d 1166 (N.D. Cal. 2007).

Controlled experiments are often misperceived as the pinnacle of scientific evidence, with all other forms of testing ranked as substandard in some way. Experiments *are* a useful way to collect evidence, but many important questions cannot be addressed through experiment. Some aspects of the natural world aren't manipulable (e.g., distant stars), and hence can't be studied with direct experiments. In other cases, it would be unethical to perform a direct experiment on the study system—for example, randomly assigning some pregnant people to receive Bendectin to assess its safety or intentionally introducing a pollutant to a sensitive ecosystem to determine its effect.

Nonexperimental methods, which involve no intentional manipulation of condition, are often critical in gathering evidence to evaluate important hypotheses. These methods still allow us to generate predictions from a hypothesis and make observations to test those predictions. For example, we can't experiment on stars to test hypotheses about which elements occur within them, but we *can* test those ideas by figuring out what wavelengths of light the presence of different elements would generate and monitoring the stars' emission spectra.[92]

Nonexperimental research describes or documents a phenomenon. We might imagine a biologist sketching a bacterium viewed through a microscope, but of course, complex machinery, custom-designed instruments, and detailed measurements may also be involved. Analyses to test drinking water for contaminants, detect radiation coming from deep space, or measure consumer perception of a particular brand (e.g., surveys to support trademark litigation) fall into this category. Often nonexperimental studies involve making some sort of comparison among observations to detect patterns and associations. For example, a key study supporting the CFC/ozone link documented ozone measurements over Antarctica and compared results obtained between 1957 and 1984, showing a pattern of dramatic decline.[93] This key piece of convincing evidence was not obtained through an experiment, but through observation and comparison over time. Similarly, many nonexperimental epidemiologic studies informed the *Daubert* litigation, some that compared birth outcomes for mothers who had taken Bendectin to those who did not (cohort studies) and some that compared Bendectin exposure between those with a birth defect and those without (case-control studies).

Some effects we might wish to study experimentally cannot be examined in that context because of ethical or logistical considerations. However, scientists can leverage situations in which such changes occur naturally, without an intentional manipulation on the part of the researcher, to gather relevant data in a nonexperimental setting. For example, studying the connection between years

92. For example, see Noriyuko Matsunaga et al., *Identification of Absorption Lines of Heavy Metals in the Wavelength Range 0.97-1.32 μm*, 246 Astrophysical J. Supp. Series 10 (2020), https://doi.org/10.3847/1538-4365/ab5c25.

93. Farman et al. 1985, *supra* note 8.

of parental schooling and children's well-being presents nearly insurmountable experimental challenges. However, researchers have taken advantage of variation in the rollout of government educational programs and policies to find comparison groups that mimic the desired experimental setup, contrasting outcomes from children whose parents received extra years of schooling with those from closely matched communities that received the program later.[94] More recently, in another comparative, nonexperimental study, researchers used data from school districts that lifted mask mandates at different times to examine the on-the-ground effect of masking on transmission of SARS-CoV-2.[95]

Many explanations can be tested through both experimental and nonexperimental methods. Each offers advantages. A laboratory experiment may allow a researcher to control many factors and make a strong case for causality; however, such experiments may not be able to closely mimic the complex environment outside the laboratory and so are vulnerable to the argument that the causal mechanism at work in the experimental setup is not important where it matters most—in the messy real-world conditions found in places like hospitals, schools, and natural ecosystems. On the other hand, a nonexperimental, comparative study of a field setting may be clearly grounded in those real-world conditions but leave open questions about confounding causal factors. For example, medical studies using nonhuman animal models may provide convincing evidence of the carcinogenic effects of a substance in the model organism but not necessarily people, while epidemiologic evidence of cancer rates in people are clearly relevant to human cancers but may be consistent with multiple hypotheses as to the cause of the cancer (see the reference guides on toxicology and epidemiology, in this manual, for further considerations). The nonexperimental masking study referenced above addressed some complaints about lab-based mask research not reflecting real-world variation in the types of masks worn in schools and the consistency with which they are worn, but the study was nevertheless criticized by opponents for not being a randomized controlled trial.[96]

Whenever possible, triangulating across a body of evidence based on multiple methodologies is the best approach to evaluating the scientific support for a hypothesis. The scientific community does not fully exclude consideration of a particular type of evidence relevant to a hypothesis unless it is clearly compromised methodologically or there are other a priori reasons for thinking it

94. Shin-Y Chou et al., *Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan*, 2 Am. Econ. J.: Applied Econ. 63–91 (2010), https://doi.org/10.1257/app.2.1.33. For more on the complex statistical models that can be used to analyze data from studies such as this one, see David H. Kaye and Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

95. Tori L. Cowger et al., *Lifting Universal Masking in Schools—Covid-19 Incidence among Students and Staff*, 387 New Eng. J. Med. 1935–46 (2022), https://doi.org/10.1056/NEJMoa2211029.

96. Roni Caryn Rabin, *Masks Cut Covid Spread in Schools, Study Finds*, N.Y. Times (Nov. 10, 2022), https://perma.cc/JVJ5-LUTX.

irrelevant. This same issue, assessing the value of the many lines of evidence relevant to a particular question (e.g., if and how nonexperimental epidemiologic studies and experimental animal studies should inform deliberation in toxic tort litigation), has presented challenges for the courts and led to decisions about admissibility that have generated controversy.[97] In science, the available evidence (some of which may come from other research programs not designed to test the hypothesis under consideration) is evaluated as a body, along with the strengths, weaknesses, and caveats relating to each type of data, an approach which, some scholars have argued, the judiciary has not always followed.[98] When a particular experimental or nonexperimental methodology has been applied to a question multiple times, special analytic techniques like meta-analyses may be used to interpret this body of evidence (see section titled "Systematic Reviews and Meta-analyses" below).

Within scientific methodologies, blinding generally refers to concealment of comparison group membership so that this knowledge does not bias the study outcome. In a double-blind clinical trial, neither the researcher nor the participants know which group is receiving a placebo and which is receiving the experimental intervention until after data collection is complete. Triple blinding refers to additionally concealing this information from the data analysts. In noncomparative studies, blinding may refer to concealment of the purpose or sponsor of the research from participants and those implementing the protocol (see the *Reference Guide on Survey Research*, in this manual). Blinding can be helpful in many investigations, experimental or not, in which human perception might tip the outcome in a particular direction—for example, in handwriting analysis and in proficiency testing to determine a lab's ability to perform a particular

97. For an overview, see Margaret A. Berger, *What Has a Decade of* Daubert *Wrought?*, 95 Supp. 1 Am. J. Pub. Health s59–s65 (2005) [hereinafter Berger 2005]. For a small sample of discussion relevant to this issue, see Gary Edmond & David Mercer, *Litigation Life: Law-Science Knowledge Construction in (Bendectin) Mass Toxic Tort Litigation*, 30 Soc. Stud. Sci. 265–316 (2000); Victoria Sutton & Brie DeBusk Sherwin, *Toxicological Animal Studies Disparate Treatment as Scientific Evidence*, 2 J. Animal & Env't L. (2011), https://perma.cc/V2U3-SXKV; and Raymond Richard Neutra, Carl F. Cranor, & David Gee, *The Use and Misuse of Bradford Hill in U.S. Tort Law*, 58 Jurimetrics 127–62 (2018) [hereinafter Neutra et al. 2018]. For argument on the opposing side, see, for example, Dije Ndreu Comment, *Keeping Bad Science Out of the Courtroom: Why Post-Daubert Courts Are Correct in Excluding Opinions Based on Animal Studies from Birth-Defects Cases*, 36 Golden Gate U. L. Rev. 459 (2006); and Amanda Hungerford Note, *Back to Basics: Courts' Treatment of Agency Animal Studies after* Daubert*,* 110 Colum. L. Rev. 70–113 (2010).

98. Some scholars have raised concerns that the courts have on occasion unfairly dismissed numerous individual lines of evidence as being flawed or insufficiently conclusive and concluded that evidence is lacking, when in fact the *body* of evidence, taken as a whole, points to a clear conclusion. For more, see discussion of *Milward v. Acuity Specialty Products Group, Inc.*; see also Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual; Berger 2005, *supra* note 97; and Steve C. Gold, *A Fitting Vision of Science for the Courtroom*, 3 Wake Forest J.L. & Pol'y 1 (2013).

technique (see the *Reference Guide on Forensic Feature Comparison Evidence*, in this manual).

Qualitative Data, Quantitative Data, and Mixed Methods

Quantitative data are numeric; qualitative data are nonnumeric, taking the form of, for example, written descriptions, audio recordings, and video footage. Both are essential forms of evidence within science, just as they are in a trial. Qualitative data are often misperceived as soft, less useful than quantitative data, and the domain of the social sciences. This view overlooks many applications of qualitative data within the natural sciences—for example, to visually identify biological species, minerals, and strata. Neither are quantitative data absent from the social sciences, as illustrated by numerous examples in the *Reference Guide on Survey Research*, in this manual. Furthermore, nuanced qualitative data from social science, psychology, and behavioral science are particularly valuable for providing context and meaning for quantitative results, as well as for suggesting hypotheses for further investigation. For some analyses, qualitative data may be validly aggregated and converted to quantitative data through a coding scheme. Mixed-methods research uses a combination of qualitative and quantitative techniques within a single investigation to detect patterns and form and test hypotheses.

When research is quantitative in nature, scientists often use common numeric standards to evaluate the outcomes of their tests:

- *Evaluating hypotheses:* In Bayesian approaches, scientists attempt to calculate the probability that a hypothesis is true given the data available. Such approaches begin by choosing a *prior probability distribution* ("priors") over mutually exclusive hypotheses. These priors are based on previous research, background theories, or even statistical indifference between possibilities. Then as observations are made or experiments conducted, a new probability distribution called the *posterior* distribution is calculated using Bayes's theorem. Over repeated observations or experiments, the values of the posterior distribution will converge to the true values. Thus, Bayesian approaches allow scientists to calculate the probability that the hypothesis is true given the evidence they have observed. In contrast, in the commonly used *p*-value approach, scientists compare a test hypothesis (e.g., that drug X is effective) to a null (e.g., that there is no difference in cure rates between those who took drug X and those who took a placebo). Scientists then calculate the probability that the null hypothesis could be true even with the observed difference between conditions (e.g., the cure rate of patients taking drug X compared to that of those taking a placebo). For more details, see the *Reference Guide on Statistics and Research Methods*, in this manual. In these cases, scientists

often use a probability of 0.05 as the threshold of significance, indicating that the null hypothesis would generate a result as or more extreme than the observed data only 5% of the time, providing support to the non-null hypothesis. For example, in Zoloft product liability litigation, the court excluded expert testimony (arguing that Zoloft caused birth defects in the children of mothers taking the drug) because the testimony was based on trends—patterns in data that do not reach the level of statistical significance—leaving a strong possibility that any association between Zoloft and birth defects was due to chance alone (the null hypothesis).[99] When considering testimony and evidence that relies on *p*-values, it is important to keep in mind a few caveats:

(1) A *p*-value lower than 0.05 does not *prove* that a null hypothesis is false. It is strong evidence, but there is a small chance that the difference observed could be the result of chance alone.

(2) Using a low *p*-value (e.g., 0.05) as a criterion for significance sets a high bar for rejecting the null hypothesis, minimizing the chance of getting a false positive—as a hypothetical example, minimizing the chance that we conclude that Zoloft use in pregnancy leads to more birth defects when, in fact, there is no relationship between Zoloft and birth defects. When the *p*-value is higher than the cutoff criterion and the null is not rejected (and note that the lower our cutoff criterion, the more likely this is to happen), it might be tempting to conclude that the null hypothesis must be true (e.g., that there is no relationship between Zoloft and birth defects). But this is erroneous. Having not enough evidence to establish a relationship is not the same as having strong evidence that there is no relationship. We should also be concerned about false negatives—that is, concluding that Zoloft use in pregnancy has no impact on birth defects in the hypothetical case that Zoloft use *does* in fact lead to more birth defects—which could have drastic implications for toxic tort litigation. The likelihood of false negatives can be assessed through an examination of statistical power (see paragraph (3) below).

(3) Depending on its design and the underlying system, a study may have limited ability to reject a null hypothesis (i.e., detect a difference) at the significance cutoff criterion. Power refers to a test's ability to reject a hypothesis that is indeed false. A study with high statistical power minimizes the chance of false negatives. Statistical power increases with the sample size of the study and with the effect size one wishes to detect (see Evaluating impact bulleted paragraph below). Well-designed studies have sufficient power to detect the differences of interest, but it may not be apparent when a test lacks power (see the *Reference Guide*

99. *In re* Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig., 26 F. Supp. 3d 449 (E.D. Pa. 2014).

on Statistics and Research Methods, in this manual, for more details), and it is possible to leverage underpowered studies to argue for no causal effect when one does in fact exist.[100]

(4) When several separate studies find nonsignificant trends, meta-analyses (see section titled "Systematic Reviews and Meta-analyses" below) may be used to examine what the data taken as a whole say, as any individual test may have limited statistical power for the reasons described above.

(5) The 5% cutoff value for significance is a convention, not a number that emerges out of nature or statistical theory, and scientists have their own debates about how *p*-values should be used.[101] There is nothing special about 0.05 other than the fact that it is an a priori criterion; clearly 0.04 or 0.06 are not very different. As described in the section titled "Assessment of Evidence" below, scientists may take into account the outcome of accepting or rejecting a particular hypothesis and adjust the design of their tests, or their threshold for action or acceptance, to guard against harm.

- *Evaluating impact:* While *p*-values and Bayesian approaches attempt to answer the question "Is there a relationship between variables?" effect sizes answer the question "How strong is this relationship?" *P*-values can generally be used to interpret the significance of a relationship in the same way across studies, but effect size can be communicated using many different statistics. Effect sizes are an important consideration at trial because they provide an indication of how much of an outcome can be attributed to a particular causal agent. For example, congenital cardiovascular defects in neonates can be caused by a variety of factors and occur relatively frequently even when mothers did not take an antidepressant during pregnancy. In the Zoloft case mentioned above, one of the studies cited by the memorandum opinion found that a different antidepressant, fluoxetine, had an adjusted odds ratio of 4.47, meaning that mothers who took fluoxetine during early pregnancy were estimated to be 4.47 times more likely to have a baby with a congenital cardiovascular anomaly than mothers who did not take the antidepressant during early pregnancy.[102] Causal relationships with a weak effect are difficult to detect without very large

100. See Neutra et al. 2018, *supra* note 97, for discussion of the judiciary's awareness of power when evaluating statistical significance.

101. For example, see Daniel J. Benjamin et al., *Redefine Statistical Significance,* 2 Nature Hum. Behav. 6–10 (2018), https://doi.org/10.1038/s41562-017-0189-z; John P.A. Ioannidis, *The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance,* 321 JAMA 2067–68 (2019), https://doi.org/10.1001/jama.2019.4582; Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar, *Moving to a World Beyond "p < 0.05"*, 73sup1 Am. Statistician 1–19 (2019), https://doi.org/10.1080/00031305.2019.1583913.

102. Orna Diav-Citrin et al., *Paroxetine and Fluoxetine in Pregnancy: A Prospective Multicentre, Controlled, Observational Study*, 66 Brit. J. Clinical Pharmacology, 695–705 (2008), https://doi.org/10.1111/j.1365-2125.2008.03261.x.

sample sizes. In the Zoloft litigation, the plaintiffs' expert's testimony would have argued that multiple, nonsignificant associations between Zoloft use and birth defects indicated a causal relationship. The testimony was excluded because these results were consistent with a weak causal relationship (a small effect size), one that is "so weak that one cannot conclude that the risk is greater than that seen in the general population."[103]

- *Evaluating estimates:* In science (and in contrast to their lay meanings), the terms *uncertainty* and *error* refer to the variability of a set of data that is intended to estimate a single number. Uncertainty and error are generally expressed as a range, within which we are confident that, if the study were repeated, the new result would fall. Scientists often use a 95% confidence interval for this purpose. For example, the researchers who first documented the ozone hole compiled daily measurements of total ozone over Halley Bay, Antarctica, taken over more than two decades. The measurements varied substantially from one day to the next. To communicate the overall pattern, the researchers estimated the maximum error bounds on their measurements and showed that the trend of decreasing ozone levels over time was much larger than any potential errors in the measurements. No matter where within the error range the true ozone values fell, the pattern they'd identified would still be exhibited.

Note that the 95% and 5% cutoffs are somewhat arbitrary, and a higher degree of confidence might be required if more certainty were desired—for example if an impactful policy decision depended on the conclusion. Similar, cross-disciplinary, standardized cutoff values for effect sizes do not exist because of the many different statistics used to communicate effect size and because interpreting these statistics is highly dependent on the system under consideration and for what purposes one needs to interpret associations. For more on these statistics, see the *Reference Guide on Statistics and Research Methods*, in this manual.

Importantly, a low *p*-value, large effect size, or narrow confidence interval found in a particular study should not be taken to indicate scientific consensus. These statistical measures help communicate the results of an isolated investigation, and scientific consensus is built over the course of many investigations and analyses (see section titled "Achieving Scientific Consensus" below).

Modeling

The term *model* is used in many different ways in science. However, as a research method of modern science, modeling almost always refers to developing and testing

103. Memorandum Opinion 9, *In Re* Zoloft Products Liability Litigation, No. 2:12-MD-2342 at 9 (E.D. Pa. 2014), ECF No. 979.

a mathematical model. These models can be relatively simple and thus analytically tractable; however, other models are so complex that they are built as a piece of computer code. One can think of a model as a surrogate, a simplified representation of a real-world system.[104] That system could be any part of the natural world, from how gases move in the atmosphere to how social media influencers arise.[105]

Models generally accept multiple different input values, or parameters, and then generate a set of predictions, which are usually, though not always, quantitative in nature. For example, models of Earth's atmosphere and climate[106] accept input about the shape of terrain, vegetation cover, ice cover, atmospheric makeup, and many other factors. The models integrate mathematical rules that simulate how gases circulate over the surface of earth, how sunlight is reflected from various surfaces, how water moves from land surfaces into the atmosphere, and more. And they generate predictions about future states of the Earth system: temperature ranges, precipitation levels, frequency of extreme weather events, etc. Essentially, a model presents an argument: If these input values represent the starting point of the system, and if the system works according to the rules implemented in the model's code or equations, then we'd expect the system to have this predicted state in the future.

Models can and must be tested against evidence, just like hypotheses. Their predictions can be compared to observations to see if the model seems to be a fair representation of the system. If a model's predictions do not match key elements of our observations or evidence, we must reevaluate our input parameters and/or the sufficiency and accuracy of the model.[107] The atmospheric models that were used to test Rowland and Molina's hypothesis incorporated many different units of scientific knowledge, which were themselves supported by other lines of evidence. While the models were not perfect representations of the Earth system, they were close enough to generate many predictions that held true of the atmosphere in general, giving scientists confidence that the models were useful approximations of actual mechanisms. And when Rowland and Molina's hypothesis was incorporated into those models, they generated two key predictions that were borne out in observations: that CFCs should be abundant in the lower atmosphere but rare in the upper atmosphere, a prediction that was immediately verified with evidence from sensors placed on aircraft and balloons,

104. Michael Weisberg, Simulation and Similarity: Using Models to Understand the World (2013) [hereinafter Weisberg 2013].

105. For example, see Penelope Maher et al., *Model Hierarchies for Understanding Atmospheric Circulation*, 57 Rev. Geophysics 250–80 (2019), https://doi.org/10.1029/2018RG000607; and Nicolò Pagan et al., *A Meritocratic Network Formation Model for the Rise of Social Media Influencers*, 12 Nature Commc'ns 6865 (2021), https://doi.org/10.1038/s41467-021-27089-8.

106. For example, see June-Yi Lee et al., *Future Global Climate: Scenario-Based Projections and Near-Term Information*, *in* Climate Change 2021: The Physical Science Basis 553–672 (2021), https://doi.org/10.1017/9781009157896.006.

107. Weisberg 2013, *supra* note 104.

and that the ozone layer would become thinner over time, a prediction that was verified many years later.[108]

If a model is supported and seems to be a good representation of the target system, we can use it to answer "what if" questions: What would have happened in 50 years if we had continued CFC production at 1974 rates? How would changing a social network's recommendation system affect influencers' ability to spread information? If a hazardous chemical is used in a continuous action air freshener, what dose would an individual confined mostly to the home inhale daily? If a business had continued to operate instead of being barred from doing so, what profits and value would it have accrued during that period of closure? Predictions from widely accepted models may be considered evidentiary and used to build other arguments—for example, about the cause of a disease or injury. The *Reference Guide on Exposure Science*, in this manual, explains the types of models that are used to estimate the exposures to and environmental fates of chemicals, and the *Reference Guide on the Estimation of Economic Damages*, in this manual, explains the types of financial models used to estimate damages.

Correlation and Causation

While the often-stated maxim that correlation does not imply causation is true, in fact, correlation is the only means that we have of establishing causation in science.[109] The reason we accept that smoking causes lung cancer is a series of very convincing correlations: between increases in population-level cigarette consumption and lung cancer rates; between exposure of lab animals to cigarette smoke tars and the development of tumors; between smoking and the presence of precancerous cells in the lungs; etc.[110] Experiments are valuable for establishing causation because they allow us to rule out many confounding variables, but interpreting experimental evidence *still* involves looking for correlations between experimental manipulations and outcomes. When many correlations line up in different tests, all linked by a logical natural explanation, scientists are likely to accept a causal relationship.[111] In our ozone layer example, when correlations were shown between the presence of ice crystals and more rapid ozone-depleting

108. For example, see A. L. Schmeltekopf et al., *Measurements of Stratospheric $CFCl_3$, CF_2Cl_2, and N_2O,* 2 Geophysical Rsch. Letters 393–96 (1975), https://doi.org/10.1029/GL002i009p00393; and Farman et al. 1985, *supra* note 8.

109. Judea Pearl, Causality (2d ed. 2009).

110. Robert N. Proctor, *The History of the Discovery of the Cigarette-Lung Cancer Link: Evidentiary Traditions, Corporate Denial, Global Toll*, 21 Tobacco Control 87–91 (2011), https://doi.org/10.1136/tobaccocontrol-2011-050338.

111. For considerations in inferring causation in the field of epidemiology, see discussion of the Bradford-Hill criteria for causation in Steve C. Gold et al., *Reference Guide on*

chemical reactions in a lab environment, between the prevalence of icy polar clouds and the degree of ozone depletion in the stratosphere, and between the predictions of models that incorporated icy clouds and observations of the atmosphere, scientists accepted that polar stratospheric clouds were part of the cause of the extreme ozone loss they observed in Antarctica.[112] Correlations are the basis of essential evidence both in science and at trial.

Assumptions and Auxiliary Hypotheses

Assumptions are often perceived as a liability in logically reaching a sound conclusion, but in fact *all* scientific tests depend on making assumptions. Even testing a simple hypothesis about the presence of lead in drinking water involves making many assumptions about how the samples were collected, how the analytic equipment was calibrated, and of course, the accuracy of the stores of scientific knowledge that go into the field of optical emission spectroscopy, which is used to determine the elemental constituents of metals. One might wonder about trusting *any* scientific evidence if it all relies on assumptions. In science, assumptions are treated as auxiliary hypotheses, which can themselves be tested, a process that is part of the fabric of scientific knowledge building. By testing hypotheses and auxiliary hypotheses in different combinations, using different methods, science can triangulate between lines of evidence, homing in on the accuracy of a particular idea, independent of the assumptions that underlie its tests.[113]

For example, the sorts of atmospheric models that were used to examine the link between CFCs and ozone depletion are complex and themselves represent hundreds of auxiliary hypotheses—about how molecules move and diffuse through the atmosphere, about how solar radiation penetrates the atmosphere, about the strength of solar radiation throughout the day, and so forth—each of which had some justification and/or evidence supporting it.[114] When Molina and Rowland discovered that chlorine nitrate was longer lived than previously reported and so could interact with chlorine atoms from CFCs, they published this information so that it could be added to atmospheric models.[115] Several groups were involved in testing the Rowland-Molina hypothesis, and up to this

Epidemiology, in this manual. For an analysis of potential judicial interpretations of Bradford-Hill, see Neutra et al. 2018, *supra* note 97.

112. For example, see Susan Solomon, *Progress Towards a Quantitative Understanding of Antarctic Ozone Depletion,* 347 Nature 347–54 (1990), https://doi.org/10.1038/347347a0.

113. Carl G. Hempel, Philosophy of Natural Science (1966).

114. For an introduction to these models and examination of how they have developed over time, see Paul N. Edwards, *History of Climate Modeling*, 2 WIREs Climate Change 128–39 (2011), https://doi.org/10.1002/wcc.95.

115. Rowland et al. 1976, *supra* note 41.

point, their models' predictions had largely agreed. But once the researchers added the chlorine nitrate reactions to their models, different groups got conflicting results. Some models even predicted an increase in ozone. The discrepancy was traced to a particular auxiliary hypothesis. Those models predicting a net increase in ozone had one thing in common: They all used the approximation that the sun shines at its average intensity all the time, instead of varying throughout the day.[116] Correcting this simplifying auxiliary hypothesis brought these models into agreement with the other models and with the actual atmospheric measurements of chlorine nitrate. Incorrect assumptions can be identified as such because of the iterative nature of the process of science.

Replication and the "Replication Crisis"

Scientists aim for their tests to be replicable—so that, for example, two studies examining consumers' perception of the "Thermos" brand name would yield concordant results if carried out in similar contexts.[117] The goal of replicability comes with science's job of uncovering the mechanisms by which the natural world operates; these mechanisms are *consistently* at work, and so, if a study's finding cannot be replicated, it suggests that our understanding of the mechanism underlying the result or our testing methodologies are insufficient. In practice, many studies are not repeated step by step in an exact replication of the original, and those that are replicated exactly may not be published.[118] Instead, hypotheses are often tested in slightly different contexts using varied methods, a process that helps establish the accuracy of the idea and further illuminates its scope and applicability. In our ozone layer example, scientists' use of multiple atmospheric models, some of which made simplifying hypotheses about the strength of solar radiation throughout the day and some of which did not, led to a better understanding of ozone depletion and of the sensitivity of the models, which are used for other purposes, like weather and climate predictions.

In recent years, concerns have arisen that many published scientific studies are unreplicable and that the hypotheses they purport to support may in fact be incorrect.[119] This concern is often termed "the replication crisis," and we will

116. F. S. Rowland, John E. Spencer, & Mario J. Molina, *Estimated Relative Abundance of Chlorine Nitrate among Stratospheric Chlorine Compounds,* 80 J. Phys. Chem. 2713–15 (1976), https://doi.org/10.1021/j100565a020.

117. In different disciplines and contexts, the terms *replicability, reproducibility*, and *repeatability* are used in inconsistent ways. Here, we use the term *replication* to mean a study carried out with the intent of mimicking another to the closest degree possible.

118. *See* NASEM, *supra* note 46, *and* Fiona Fidler & John Wilcox, *Reproducibility of Scientific Results*, *in* The Stanford Encyclopedia of Philosophy (Edward N. Zalta ed., 2021).

119. *See* NASEM, *supra* note 46, and more recent publications such as Timothy M. Errington et al., *Investigating the Replicability of Preclinical Cancer Biology*, 10 eLife e71601 (2021), https://doi.org/10.7554/eLife.71601.

use that terminology for consistency.[120] The identification of many unreplicable studies leads to valid questions about science's reward structure of promoting low quality and misleading practices. Individual scientists and scientific institutions have responded with proposals for incentivizing exact replication studies, incentivizing the reporting of negative or nonsignificant results, encouraging transparent reporting, diversifying peer review, and changing aspects of science's reward structure that seem to be to blame for the apparent proliferation of unreplicable studies.[121] Concrete steps aimed at achieving these goals have been taken by, for example, expanding the use of platforms where researchers can preregister a planned investigation, describing hypotheses to be examined, methods used, and planned analyses. Preregistration may help reduce the likelihood that researchers will fail to report null results or mine their data for significant results, which will occur by chance alone 5% of the time with a p-value of 0.05, as described above. Preregistration is standard practice for medical clinical trials, but it is becoming more common, especially within psychology. However, it remains to be seen to what extent, if any, preregistration will improve replicability; research on this question is still in progress.[122] Nevertheless, preregistration is just one of many initiatives that scientific institutions are pursuing to help address the replication crisis, reflecting their commitment to encouraging high-quality research.[123]

While there is no question that changing science to promote higher-quality research is worthwhile and will make science more efficient at building accurate knowledge about the natural world, the fact that there is room for improvement in the process of science does not necessitate distrust of hypotheses that have gained widespread acceptance in the scientific community and about which consensus has been achieved. Indeed, some have argued that the problem of unreplicated results may not be any worse today than it has been over the history

120. However, as discussed herein, describing the situation as a "crisis" implies that it is a new situation and a catastrophic problem for scientific progress, which does not seem to be the case.

121. For examples of proposed changes, see Marcus R. Munafò et al., *A Manifesto for Reproducible Science*, 0021 Nature Hum. Behav. 1–9 (2017), https://doi.org/10.1038/s41562-016-0021; and Piers D. L. Howe & Amy Perfors, *An Argument for How (and Why) to Incentivize Replication*, 41 Behav. & Brain Scis. e135 (2018), https://doi.org/10.1017/S0140525X18000705, as well as NASEM, *supra* note 46.

122. For some research on this question, see A. Claesen, S. Gomes, F. Tuerlinckx, & W. Vanpaemel, *Comparing Dream to Reality: An Assessment of Adherence of the First Generation of Preregistered Studies*, R. Soc. Open Sci. 211037 (2021), http://doi.org/10.1098/rsos.211037, and Christopher Allen & David M. A. Mehler, *Open Science Challenges, Benefits and Tips in Early Career and Beyond*, 17 PLoS Biology e3000246 (2019), https://doi.org/10.1371/journal.pbio.3000587. For discussion of why preregistration might not work as planned, see Kai Kupferschmidt, *More and More Scientists Are Preregistering Their Studies. Should You?*, Science (2018), https://doi.org/10.1126/science.aav4786.

123. Max Korbmacher et al., *The Replication Crisis Has Led to Positive Structural, Procedural, and Community Changes*, 1 Commc'ns Psych. 1–13 (2023), https://doi.org/10.1038/s44271-023-00003-2.

of Western science and that, in any case, this problem is not impeding progress at building new, accurate knowledge about the natural world.[124]

The replication crisis is a concern about individual investigations, not scientific consensus. Not all published ideas will be replicated, but neither are they all worth trying to replicate. Hypotheses that seem fruitful, explanatory, and potentially important are useful for understanding the world and so will be tested in different ways, by different people. In science, the fact that a particular study has not been precisely replicated does not mean it is invalid. Nor is a successful replication attempt necessarily an indication that the idea being tested is correct. Only ideas that have held up to repeated testing in different ways, having built up a body of supporting evidence, are likely to gain wide acceptance within science. However, judges are frequently asked to consider evidence that is far from achieving consensus. The replication crisis makes clear that peer review alone cannot ensure that the conclusions of published studies are actually correct, highlighting the responsibility judges bear in evaluating the validity of the methodologies that contributed to a particular piece of research.

Achieving Scientific Consensus

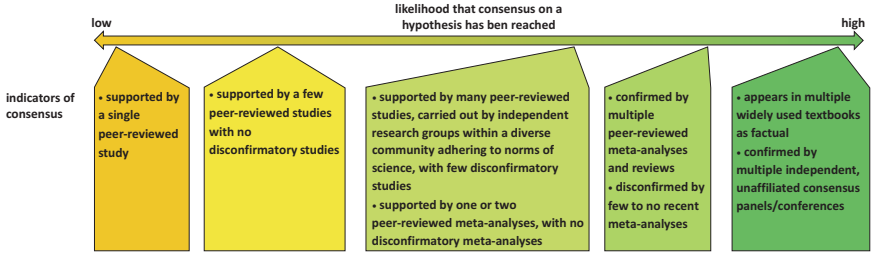
Much of the preceding text, as well as subsequent reference guides, will assist judges in evaluating the methods behind evidence relating to emerging or contentious science. Here, we examine the opposite end of the continuum of certainty: scientific consensus or widespread acceptance, the fifth factor that *Daubert* guides judges toward in their consideration of the reliability of expert testimony—though we emphasize that consensus is not a requirement for admissibility. While widespread acceptance provides a strong indicator of the reliability of scientifically acquired knowledge, there is no surefire way of assessing which hypotheses have reached this level of acceptance. No regular polls of scientists assess scientific consensus, although on some issues with significant societal importance and controversy, like climate change, such polls may be carried out.[125] Sometimes consensus conferences or panels are convened or consensus reports are written to assess the state of the field and/or to better communicate important conclusions to the public and policy makers. For example, several organizations, such as the

124. *See* Daniele Fanelli, *Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?*, 115 PNAS 2628–31 (2018), https://doi.org/10.1073/pns.1708272114; *and* Richard M. Shiffrin, Katy Börner, & Stephen M. Stigler, *Scientific Progress Despite Irreproducibility: A Seeming Paradox*, 115 PNAS 2632–39 (2018), https://doi.org/10.1073/pnas.1711786114.

125. For examples, see Peter T. Doran & Maggie Kendall Zimmerman, *Examining the Scientific Consensus on Climate Change*, 90 Eos 22–23 (2011), https://doi.org/10.1029/2009EO030002; Neil Stenhouse et al., *Meteorologists' Views About Global Warming: A Survey of American Meteorological Society Professional Members*, 95 Bull. Am. Meteorological Soc'y 1029–40 (2014), https://doi.org/10.1175/BAMS-D-13-00091.1.

Community Preventive Services Task Force and the Cochrane Collaboration, focus on producing reports that assess scientific consensus on medical questions, and the Intergovernmental Panel on Climate Change has, since 1990, produced reports assessing the scientific consensus on climate change. Similarly, the National Research Council and National Academies regularly produce consensus reports on a variety of scientific issues, including topics as diverse as cryptography, policing outcomes, and the health effects of electric and magnetic fields.[126] Returning to our ozone example, in 2003 the World Meteorological Association issued a report synthesizing scientific knowledge on ozone depletion to which 275 disciplinary experts contributed.[127] Such reports are valuable indicators of scientific consensus, but many scientific topics in question at a trial are unlikely to have such definitive documentation of acceptance to which to turn. Figure 3 summarizes some of the signals that may indicate scientific consensus has been reached, as well as signs that insufficient evidence has accumulated for the hypothesis to reach widespread acceptance. The link between CFCs and ozone depletion began at the left side of this diagram—as a single peer-reviewed study—and over the course of 15 years, accumulated so much evidence and withstood so many rounds of scrutiny that it has now achieved widespread acceptance and the highest level of certainty science has to offer, placing it at the far right side of the diagram.

Figure 3. Indicators of scientific consensus. Scientific consensus is formed based on a body of evidence, not a single study or investigation.



126. National Research Council, Possible Health Effects of Exposure to Residential Electric and Magnetic Fields (1997), https://doi.org/10.17226/5155; NASEM, Cryptography and the Intelligence Community: The Future of Encryption (2022), https://doi.org/10.17226/26168; NASEM, Policing to Promote the Rule of Law and Protect the Population: An Evidence-Based Approach (2022), https://doi.org/10.17226/26217.

127. World Meteorological Organization, *Scientific Assessment of Ozone Depletion: 2002*, Global Ozone Research and Monitoring Project—Report No. 47 (2003).

The path to consensus generally involves the accumulation of multiple lines of converging evidence, as studies are conducted, published, scrutinized, and iterated on, bringing more and more of the community into agreement.[128] This process works most effectively when it is carried out by a diverse group of scientists adhering to the expectations and norms outlined above. Even once broad consensus has been achieved, it is unlikely to be completely unanimous. Because scientific explanations are inherently fallible and science values skepticism, it is typical for a few in the community to remain unconvinced by even extremely compelling evidence. In fact, in a state of consensus, having a few holdouts can help maintain a degree of vigilance among the persuaded, which helps ensure the reliability of accepted knowledge. When a judge is presented with a disagreement among scientific experts, it is reasonable to seek clarification on how representative of the scientific community the two views are. Is this a case of truly unsettled science, where methodologically sound studies have been carried out and scrutinized, but open questions and disagreements remain because the hypothesis has not yet been thoroughly investigated, in which case, each party may have comparable claims on scientific reliability? Or is it a case of relatively settled science, in which one party has recruited experts who interpret the evidence differently than most of the community?[129] Note that public perception of the certainty of a scientific concept or hypothesis may differ from the actual stage of consensus building within the scientific community. This sometimes occurs as a result of strategic manipulation from stakeholders who stand to be harmed if the public were to understand the true state of scientific consensus surrounding the hypothesis, as has occurred with, for example, the health effects of tobacco, ozone depletion, and climate change.[130]

Despite there being a culture of skepticism in science, consensus is often reached over time. Hypotheses are more likely to gain widespread acceptance and achieve scientific consensus if they:

128. Mary Jo Nye, Molecular Reality: A Perspective on the Scientific Work of Jean Perrin (1972).

129. For example, see *Allen v. Hyland's, Inc.*, which allowed testimony from experts in homeopathy, a set of rejected hypotheses about medical causation, and from a scientist noted for holding fringe views in the scientific community. *See* Kristin Shrader-Frechette, *Conceptual Analysis and Special-Interest Science: Toxicology and the Case of Edward Calabrese*, 177 Synthese 449–69 (2010), https://doi.org/10.1007/s11229-010-9792-5; *and* Jan Beyea, *Lessons to be Learned From a Contentious Challenge to Mainstream Radiobiological Science (the Linear No-Threshold Theory of Genetic Mutations),* 154 Env't Res. 362–79 (2017), https://doi.org/10.1016/j.envres.2017.01.032.

130. For example, see Michael E. Mann, *The New Climate War: The Fight to Take Back our Planet* (2021); Naomi Oreskes, *The Scientific Consensus on Climate Change*, 306 Science 1686 (2004), https://doi.org/10.1126/science.1103618; Oreskes & Conway 2010, *supra* note 2; and G. Supran, S. Rahmsdorf & N. Oreskes, *Assessing ExxonMobil's Global Warming Projections*, Science 153-162 (2023), https://doi.org/10.1126/science.abk0063.

- have been tested and garnered support many times in many ways,
- help explain disparate and previously unexplained observations,
- can more closely explain our observations than other hypotheses,
- can be broadly applied, and
- are consistent with well-established hypotheses in neighboring fields.

Systematic Reviews and Meta-analyses

Two types of publication can be particularly useful in understanding and assessing the lines of evidence relevant to a hypothesis across multiple investigations. Systematic reviews attempt to synthesize and summarize the current state of research on a topic. In them, the authors delineate both the criteria that studies must meet for inclusion in the review and the methods that will be used to assess the studies. Meta-analyses extend this process by using statistics to analyze data from multiple studies.[131] Systematic reviews and meta-analyses are used across the broad array of topics that science seeks to understand—from that which can be tightly controlled (e.g., a meta-analysis of randomized controlled trials of antimanic treatments) to complex and confounded social issues (e.g., whether school-choice programs boost student achievement).[132] For example, in the years after Rowland and Molina published their hypothesis, scientists worked to understand how the depletion of the ozone layer might affect life on Earth. One meta-analysis examined 62 field-based studies designed to assess how an increase in ultraviolet-B radiation, which the ozone layer helps filter out, might affect plant life, concluding that the effects on plants were likely to be subtle.[133] Meta-analyses are an important line of scientific evidence for the courts, but, as with all scientific methodologies, require the rigorous application of reliable methodologies, as recognized in *In re Paoli R.R. Yard PCB Litigation*, in which the exclusion of a contested meta-analysis was overturned.[134]

Assessing the state of scientific consensus often requires some degree of subject matter expertise and evaluation, which could be vulnerable to personal biases. Meta-analyses do not entirely sidestep these criticisms, as such analyses

131. Many guidelines for systematic reviews and meta-analyses appear in the peer-reviewed literature; however, these are generally oriented toward specific disciplines, making it difficult to cite a publication that applies broadly across the sciences.

132. Aysegül Yildiz et al., *Efficacy of Antimanic Treatments: Meta-Analysis of Randomized, Controlled Trials*, 36 Neuropsychopharmacology 375 (2011), https://doi.org/10.1038/npp.2010.192; Huriya Jabbar et al., *The Competitive Effects of School Choice on Student Achievement: A Systematic Review*, 36 Educ. Pol'y 247–81 (2022), https://doi.org/10.1177/0895904819874756.

133. Peter S. Searles, Stephen D. Flint, & Martyn M. Caldwell, *A Meta-Analysis of Plant Field Studies Simulating Stratospheric Ozone Depletion*, 127 Oecologia 1–10 (2001), https://doi.org/10.1007/s004420000592.

134. 916 F.2d 829 (3d Cir. 1990).

are impacted by an array of assumptions and methodological approaches, about which practitioners may disagree.[135] For this and other reasons, some sociologists of science have explored approaches to assessing consensus that examine emergent properties of consensus building and that can be applied without an evaluation of the subject matter and evidence at hand, and without polling scientists or convening expert panels. For example, Shwed and Bearman examined patterns of citations within the scientific literature to assess the timing of consensus formation around the hypotheses that smoking causes cancer, that coffee does not cause cancer, and that vaccines do not cause autism.[136] This approach has also been leveraged to assess scientific consensus on the question of whether children raised by same-sex parents have similar outcomes to children raised in other family settings,[137] an analysis motivated by legal debate on the issue.[138]

Not Necessarily Consensus

A few, easy-to-grasp (and so, tempting) criteria should not be taken as indicators of scientific consensus that a hypothesis is correct. These deceptive heuristics include:

- *Being called a law:* The terminology used to refer to a scientific explanation—whether hypothesis, theory, law, or model—does not indicate, in and of itself, that an idea is accurate. "Mere" hypotheses may have garnered much support and may be widely accepted, while some scientific laws have been overturned or are understood to have many exceptions. For example, Lamarck's second law stated that traits acquired during an individual's lifetime (e.g., muscular strength developed by lifting weights) will be passed down to one's offspring—an idea that is at odds with most of modern genetics.
- *Statistical significance:* A statistically significant result (e.g., low p-value) or high confidence in a particular measurement does not itself indicate scientific consensus in the context of a single study. For example, a 2012 study found a small but statistically significant association between exposure to

135. Jop de Vrieze, *Meta-Analyses Were Supposed to End Scientific Debates. Often, They Only Cause More Controversy,* Science 18 (Sep. 18, 2018), https://perma.cc/H39R-XTXN.

136. Uri S. Shwed & Peter S. Bearman, *The Temporal Structure of Scientific Consensus Formation*, 75 Am. Socio. Rev. 817–40 (2010), https://doi.org/10.1177/0003122410388488.

137. jimi adams & Ryan Light, *Scientific Consensus, the Law, and Same Sex Parenting Outcomes*, 53 Soc. Sci. Rsch. 200–310 (2015), https://dx.doi.org/10.1016/j.ssresearch.2015.06.008.

138. In oral arguments in Hollingsworth v. Perry, Justice Scalia stated, "... there's considerable disagreement among sociologists as to what the consequences of raising a child in a single-sex family, whether that is harmful to the child or not." (Transcript of oral argument at 19, *Hollingsworth v. Perry*, 570 U.S. 693 (2013), https://www.supremecourt.gov/oral_arguments/argument_transcripts/2012/12-144_5if6.pdf).

Zoloft in utero and major congenital malformations.[139] But other studies have reached different conclusions on this question, and a later meta-analysis that included the 2012 study found no significantly increased risk for major congenital anomalies associated with Zoloft use during pregnancy.[140] Despite an earlier statistically significant result, scientific evidence appears to be converging around there being no or minimal elevated risk for these birth defects as a result of a pregnant person's Zoloft use. While one's confidence in the conclusions of a lone study should always be tempered, if a particular method, measure, or analytic technique has been established over the course of multiple studies and is widely accepted by the scientific community, then one might well feel highly confident about the reliability of a significant result obtained through a particular application of that approach, bearing in mind that statistically significant results necessarily occur by chance alone some of the time.

- *Peer-reviewed publication:* This is an important milestone along the road to achieving consensus, but of course, many hypotheses published in peer-reviewed journals are ultimately rejected by the scientific community and do not make it very far in that journey. For example, before Watson and Crick published their ideas about the structure of DNA, a peer-reviewed publication proposed the now-rejected hypothesis that DNA is a three-stranded molecule.[141]
- *Journal prestige, scientific eminence, or high citation metrics alone:* Hypotheses published in journals with high standards for evidence and stringent peer-review practices that then go on to receive many citations because other scientists are building on the research *is* an indication that the scientific community is moving toward broad acceptance of a hypothesis; however, any one of these factors in isolation is a poor indicator of consensus. The fact that an article has high citation metrics, was published in a prestigious journal (often loosely indicated by the journal's impact factor), or has a famous author is not by itself a reason to conclude that the study's findings are widely accepted scientific facts. The vagaries, inconsistencies, and biases of publishing are discussed above, and plenty of renowned scientists have been wrong in prestigious journals. The proposal for the three-stranded DNA helix was authored by Linus Pauling, winner of the Nobel Prize in Chemistry, in the *Proceedings of the*

139. Espen Jimenez-Solen et al., *Exposure to Selective Serotonin Reuptake Inhibitors and the Risk of Congenital Malformations: A Nationwide Cohort Study*, 2 BMJ Open e001148 (2012), https://doi.org/10.1136/bmjopen-2012-001148.

140. Shan-Yan Gao et al., *Selective Serotonin Reuptake Inhibitor Use During Early Pregnancy and Congenital Malformations: A Systematic Review and Meta-Analysis of Cohort Studies of More than 9 Million Births*, 16 BMC Med. 205 (2018), https://doi.org/10.1186/s12916-018-1193-5.

141. Linus Pauling & Robert B. Corey, *A Proposed Structure for the Nucleic Acids*, 39 Proc. of the Nat'l Acad. of Sci. USA 84–97 (1953), https://doi.org/10.1073/pnas.39.2.84.

National Academy of Sciences, for example. Furthermore, women, people of color, other historically oppressed groups, and non-Western people have been, and in many cases continue to be, unfairly dismissed and so may lack the recognition their work deserves, which includes prestigious appointments and publications.[142] Ultimately, evidence, not reputation, determines scientific acceptance, although these other factors are likely to influence how long it takes the community to reach consensus and who receives credit for developing the knowledge.

Because scientific consensus emerges from community-level processes driven by scrutiny, additional testing, and application of a hypothesis, the simple indicators listed here are not, in isolation, signals of consensus.

Myths About Science

Above we described and debunked some common myths about science, which include the following.

Myth: There is a single scientific method that all scientists follow.
Fact: The process of science is nonlinear and dynamic.

Myth: The institution of peer review ensures that all published papers are sound and dependable.
Fact: The conclusions that many peer-reviewed articles reach will ultimately turn out to be incorrect.

Myth: Without an experiment, a scientific investigation cannot be rigorous or reliable.
Fact: Nonexperimental evidence is critical in evaluating many scientific ideas.

Myth: Qualitative data are "soft" and relatively unimportant.
Fact: Qualitative data are essential to providing context and meaning to quantitative results, especially within social science, and are key to generating hypotheses in many disciplines.

A few other common misconceptions about science warrant unpacking.

Myth: Hard sciences are more rigorous and scientific than so-called soft sciences.

142. *See* Esther A. Odekunle, *Dismantling Systemic Racism in Science*, 369 Science 780–81 (2020), https://doi.org/10.1126/science.abd7531 and references therein.

Fact: Both the classic hard sciences (physics, chemistry, biology, geology, and astronomy) and the social sciences, which are sometimes portrayed as soft, can be approached with well-designed, rigorous investigations and develop reliable explanations for phenomena.[143]

As later reference guides in this manual make clear, many questions dealing with economics, decision making, and human psychology are legitimate topics of scientific investigation and employ rigorous, reliable methodologies. A variety of experimental, nonexperimental, qualitative, and quantitative methodologies are employed by all of these fields, as well as in the traditional natural sciences, and judges can apply similar considerations regarding the process of science, as outlined in this reference guide, to all of them. Of course, the more complex a study system is, the more difficult it is to control variables and untangle causal mechanisms. Furthermore, tightly controlling a study system can make it more difficult to apply those results to real-world problems. Many systems of concern in the social sciences, earth sciences, and biology are enormously complex, posing methodological challenges; however, because scientific questions in these fields are also of pressing importance to human well-being, scientists go to great lengths to meet and overcome such challenges.

Myth: Science undergoes periodic scientific revolutions, or paradigm shifts, in which an old understanding of a natural system is replaced by a new one, and the two provide such different worldviews that it is not possible to make comparisons of evidence, or perhaps even communicate, across the two perspectives.

Fact: Science moves forward incrementally, as well as in large leaps. These changes occur because evidence has accumulated, and the new view makes sense of more evidence more coherently than did the old one.

Philosopher Thomas Kuhn famously conceptualized large-scale changes in accepted scientific knowledge as paradigm shifts.[144] According to Kuhn, the history of science can be divided up into times of normal science, when scientists add to and elaborate on an accepted scientific theory, such as Newton's classical mechanics, and briefer periods of revolutionary science, when there is a switch to a new theory or paradigm—in this case, Einstein's theory of relativity, which explains everything that classical mechanics did and more. Kuhn argued that new and old paradigms were incommensurable, though later retreated from this view somewhat. Philosophers and historians of science generally agree that many details of Kuhn's concept of paradigm shifts do not align well with how changes

143. Larry V. Hedges, *How Hard Is Hard Science, How Soft Is Soft Science? The Empirical Cumulativeness of Research*, 42 Am. Psych. 443–55 (1987), https://doi.org/10.1037/0003-066X.42.5.443.

144. Thomas S. Kuhn, The Structure of Scientific Revolutions (1962).

in scientific explanations actually happened.[145] While Kuhn's ideas are not a broadly accepted account of scientific change among modern philosophers, sociologists, and historians of science, they were influential in many ways, for example by focusing attention on theory change within science and the potential role of the resolution of evidentiary anomalies in this process.

Myth: Science moves forward only by falsifying, rejecting, or disproving hypotheses, leaving the "last hypothesis standing" as the most likely to be correct.

Fact: Science can neither prove nor disprove hypotheses, and scientists evaluate hypotheses based on both supporting and refuting evidence.

This misconception is based on the idea of falsification, philosopher Karl Popper's influential account of scientific justification, which suggests that all science can do is reject, or falsify, hypotheses and that science cannot find evidence that *supports* one idea over others.[146] Falsification is a popular philosophical doctrine, especially with scientists, and apparently with judges as well, as the *Daubert* court explained, "Scientific methodology today is based on generating hypotheses and testing them to see if they can be *falsified* [emphasis added]; indeed, this methodology is what distinguishes science from other fields of human inquiry."[147] However, falsification is not a complete or accurate picture of how scientific knowledge is built. First of all, reviewing the logical arguments presented in any scientific journal will reveal that evidence can and does play a role in *supporting* particular hypotheses over others, not just in ruling some ideas out, as implied by the doctrine of falsification. Furthermore, philosophers of science now recognize that science cannot once-and-for-all, absolutely *prove* any idea to be false or true.[148] Even if all the available evidence lines up for or against a particular hypothesis, science always allows for the outside possibility that we are missing something—for example, that one of our assumptions about the system is incorrect and so our tests are not measuring what we think they are, that our interpretation of the evidence is shaded by societal norms and that future generations will see the results differently, or that there is a subtle bug in the code of a common statistical package that is leading everyone in the same wrong direction. Those are certainly very, very unlikely possibilities, but they are not strictly impossible; hence, the essential fallibility of scientific knowledge. Of course, this does not

145. Kitcher 1995, *supra* note 5.

146. Karl R. Popper, Conjectures and Refutations: The Growth of Scientific Knowledge (1963).

147. See Daubert, 509 U.S. 579 at 593 (citing Michael D. Green, Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of *Agent Orange* and Bendectin Litigation, 86 Nw. U. L. Rev. 643 (1992)).

148. Godfrey-Smith 2003, *supra* note 3; Kitcher 1983, *supra* note 27; *and* Kitcher 1995, *supra* note 5.

mean that all scientific hypotheses are equally likely; in fact, it's the opposite. The whole business of science is sorting through explanations and collecting evidence to try to find those that are most expansive and powerful and are supported by the widest variety of evidence. Despite the fact that scientists can never reach absolute proof or disproof, there are plenty of hypotheses that scientists are remarkably certain are either true or untrue, and these ideas are extraordinarily unlikely to ever be overturned. Nonscientists justifiably place their trust in this same set of reliable, accepted hypotheses every time we rely on the myriad of technologies, policies, and other applications that derive from them.

Myth: If a scientific idea is called a *hypothesis* or a *theory*, this indicates that the explanation is uncertain, whereas *laws* are nearly irrefutable scientific ideas.[149]

Fact: Hypotheses, theories, and laws are scientific explanations that differ in breadth, not in level of support, and these terms are not consistently used across scientific disciplines or throughout history.

In everyday language, the words *hypothesis* and *theory* usually refer to educated guesses or ideas that we are quite uncertain about, while *laws* are viewed as rigid mechanisms that have few exceptions. However, the meaning of these words differs within modern science, where *hypotheses* are generally understood to be potential explanations for natural phenomena regardless of the extent to which the explanations have been investigated or the amount of evidence supporting or refuting them; the term *law* is often used to refer to a statement (often a mathematical statement) about how observable phenomena are related; and *theories* are understood to be deep explanations that apply to a broad range of phenomena and that may integrate many hypotheses and laws. Any individual hypothesis, theory, or law may have accumulated strong evidence supporting or refuting it—or little in either direction. Adding further confusion, use of these terms is not standardized, and they have gained popularity in different scientific disciplines and at different points in history. Thus, we have Newton's *laws*, which are still widely applicable but were subsumed at an explanatory level by Einstein's now accepted *theory* of relativity; Lamarck's *laws*, now rejected, their explanatory position occupied by the *theory* of evolution; and the once-heretical prion *hypothesis*, now settled science and the basis for the 1997 Nobel Prize for Medicine or Physiology. In practice, judges should set little store by the terminology used to refer to a unit of purported scientific knowledge and focus instead on the methodologies and evidence that underlie it.

Myth: Scientists are completely objective in their evaluation of scientific ideas and evidence.

149. McComas 1998, *supra* note 4.

Fact: Scientists' work is influenced by their biases, motivations, backgrounds, and experiences.

Scientists may be motivated to different degrees by curiosity, by the desire to solve a problem or gain recognition or funding, or by personal financial interests. These factors and many other conscious and unconscious biases shape the questions scientists ask, how they investigate these questions, how they interpret evidence, and how they evaluate and credit the work of others. However, because scientists are expected to, incentivized to, and generally do strive for objectivity and because of the ongoing community scrutiny of scientific findings from many different viewpoints, such biases are attenuated and ultimately corrected—but this does take time. In the context of a trial concerning as-yet unsettled science, judges can reasonably weigh potential sources of bias, particularly when research has shown this to influence results, as in the case of funding bias.

Myth: Research that comes out of universities is pure and unbiased by financial and other considerations.
Fact: All research is vulnerable to bias.

The same arguments about bias that apply to individual scientists also apply to research institutions of all sorts. Universities, government agencies, organizations, and industries all have complex internal and external incentive systems that shape how they operate and carry out research. Universities may earn revenues from patents, and many have offices devoted to handling intellectual property. Such conflicts of interest can be relevant in assessing potential sources of bias at trial.

Science and the Law

Science and the law are sometimes seen to be clashing cultures, but in fact they have many overlapping properties and interests.[150] Both are oriented toward discovering truths and facts. Both are complex endeavors, involving many interacting individuals and institutions that, while guided by overarching ideals, are nonetheless embedded in and shaped by the cultures and values of society and of their institutions. Both endeavors are influenced by the backgrounds, experiences, biases, and values of their human practitioners but have mechanisms and guidelines designed to help them achieve their aims and identify truths. Here, we outline further similarities and differences between science and the law that may help judges bridge these two cultures and make useful connections between them.

150. Jasanoff 2005, *supra* note 5.

Assessment of Evidence

Both science and law rely on evidence. Judges use scientific evidence to strengthen their ability to understand the factual basis of a dispute to reach a fair and just resolution of a conflict. Judges must make preliminary assessments of scientific information presented as evidence at trial to ensure that information meets appropriate standards of scientific integrity for consideration by a jury. Scientists, on the other hand, use evidence to make judgments about the accuracy of hypotheses.

Both institutions put individuals in the position of weighing the often-incomplete evidence to come to a conclusion about what the likely truth is. In the courtroom, judges and juries must make decisions based on the current state of the scientific evidence presented during a trial. This distinct terminus to deliberation is a requirement of the law but is not generally a part of science. In most cases, scientists can and will reserve judgment on a hypothesis if warranted, waiting for more evidence before accepting or rejecting it. This difference between science and the law was acknowledged in the decision on the Zoloft litigation referenced above: "The Court recognizes that the final scientific verdict as to whether Zoloft can cause birth defects may not be delivered for many years. Nevertheless, Plaintiffs chose when to file their cases, and the Court concludes that for the Plaintiffs who have continued to pursue their claims, the litigation gates must be closed."[151]

Within the law, the standards for assessing evidence vary by jurisdiction and type of action. Criminal cases, for example, require proof beyond a reasonable doubt to draw the conclusion of wrongdoing, while civil cases require only a preponderance of evidence. As described in *The Admissibility of Expert Testimony*, in this manual, scientific evidence is admissible if the judge finds it "more likely than not that the expert's methods are reliable and that they are reliably applied to the facts at hand." Scientists too may vary their standards for evidence depending on the situation and the implications of the conclusion, but these recalibrations are not standardized across discipline or situation, as is the case in litigation. Scientists may, for example, lower their threshold for action when the investigators perceive a threat of harm, as occurred when Molina and Rowland began lobbying for policy change long before scientific consensus around the connection between CFCs and ozone depletion was achieved.

In the United States, our common-law tradition often relies on a lay jury to assess evidence to resolve disputes related to science. However, in science, only in rare cases is a specific body tasked with assessing the evidence and determining scientific consensus (see section titled "Achieving Scientific Consensus" above). Instead, consensus is usually an emergent property of interactions within the scientific community. Within science, disputes about the accuracy of

151. *In re* Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig., 176 F. Supp. 3d 483 (E.D. Pa. 2016).

hypotheses are managed by experts in that field and ultimately depend on the evidence, not the lay public's perception of the issues. For example, scientific experts have long since reached consensus on the fact that measles, mumps, and rubella (MMR) vaccines do not cause autism; however, this concern still looms large for some of the lay public, continues to make its way into the courts, and may find sympathies within the jury box.

The emphasis in the legal system on resolving legal conflicts within the common-law tradition may require presentation of scientific testimony in ways that are uncommon in the ordinary course of science. For example, the legal system allows parties or the court to select the scientific experts who will testify, advise, or mediate, and parties involved with the case may identify experts they perceive to be sympathetic to their perspective, rather than seeking out the most knowledgeable scientists or scientists whose interpretation of the evidence represents that of most of the relevant scientific community. In addition, in litigation, attorneys and judges determine how and what scientific evidence is presented, potentially limiting the jury's access to certain information. In contrast, interactions within the scientific community are not tightly regulated in terms of who is permitted to present and assess what evidence and in what ways, although journal editors and peer reviewers do play a role in determining what scientific evidence is available for review through publications. In science, debates relating to scientific evidence generally play out over many years, most notably by interactions among any subject matter experts who opt into the discussion, without imposed limits regarding what information can be used to buttress or refute an argument.

Precedent and Self-Correction

In the United States, precedent is built into our legal system. According to stare decisis, cases with the same facts are expected to be decided in the same way. Though precedent can be overturned, the hierarchical structure of the judiciary makes this occurrence rarer than it would be otherwise, and precedence remains a powerful doctrine in the law. Precedence has a limited role in science, on the other hand. Hypotheses are judged according to the specific evidence relevant to them, and scientists are expected to give up a previously accepted hypothesis if new evidence or interpretations warrant it. This is considered a normal part of good science, and science has built in mechanisms that allow for self-correction. Science is, of course, done by people, who may cling to ideas or practices, in institutions with cultures that may be slow to change, so science *does* experience a form of precedent. But because this is not a formal mechanism of science, and indeed conflicts with the idealized modes of reasoning that scientists expect of themselves, precedence is less important in science than in the law. This

difference can contribute to a lag time between a shift in scientific consensus and when this consensus is consistently applied to proceedings in courts. For example, bitemark evidence has little methodologically sound science to support its use (see the *Reference Guide on Forensic Feature Comparison Evidence*, in this manual), a fact that started to become clear in the early 2000s.[152] Yet, bitemark evidence continues to be admitted in some courts.

Conclusion

We've presented science as a human endeavor emerging out of a community that has tasked itself with working together to answer a wide variety of questions about the natural and human world through an iterative, self-correcting process. While science is often slow and its hypotheses fundamentally tentative, it is nonetheless a reliable means of generating explanations. On the broadest scale, science helps us understand the way the world is, connects disparate phenomena, and makes accurate and consistent predictions. At a practical level, science helps us solve problems, develop new technologies, make decisions, form policies, and, as this manual illustrates, administer justice. Subsequent reference guides in this manual provide a road map to the scientific findings that are most likely to factor into judicial decisions, guiding judges and juries in their search for truth and in their service to the law.

152. Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J.L. & Biosciences 538–75 (2016), https://doi.org/10.1093/jlb/lsw045.

Glossary of Terms

auxiliary hypothesis. An assumption made in the context of testing a main hypothesis.

blinding. The concealment of comparison group membership so that this knowledge does not bias the study outcome.

control group. A condition or group that does not receive the experimental manipulation, which serves to increase confidence that the change observed in the experimental group is caused by the manipulation or factor in question.

controlled experiment. An experiment that seeks to keep variables other than the experimental manipulation the same to help isolate the cause of any change.

effect size. A statistic that communicates the strength of an association between variables.

error. In reference to statistics, the potential difference between a computed or measured value and the true value.

experiment. A testing methodology that involves intentionally manipulating some factor in a system to learn how that affects the outcome.

fallibilism. The principle that holds that scientific knowledge is always open to rejection or revision, no matter how much prior evidence supports it.

falsification. A now outdated account of scientific justification, which proposes that science can only prove hypotheses wrong and that scientific acceptance arises through a process of elimination.

hypothesis. Within science, a potential explanation for a natural phenomenon, regardless of the extent to which the explanation has been investigated or the amount of evidence supporting or refuting it, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

law. Within science, an often-mathematical statement of the relationship among observable phenomena, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

meta-analysis. A study that uses statistics to analyze data from multiple other studies, aggregating and synthesizing their results.

mixed methods. A research approach that uses a combination of qualitative and quantitative techniques within a single investigation to detect patterns and to form and test hypotheses.

model. Usually, a mathematical representation of hypothesized mechanisms and interactions within a system that can be used to investigate the impact

of parameter changes on predicted system outcomes, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

natural. Concerning any element of the physical universe, whether made by humans or not.

peer review. In science, a standardized process in which journal articles (or other scientific communications, such as funding applications) are vetted by other subject matter experts before acceptance or publication.

placebo. In medical studies, a treatment, medicine, or therapy given to study participants that is not known to have a therapeutic effect on the condition of interest.

power. A measure of a statistical test's ability to detect an effect that is present.

prediction. A potential outcome of a scientific test that is arrived at by logically reasoning about what we would expect to observe if a hypothesis were true or false.

preprint. A fully drafted scientific journal article that includes all the features of a published article but that has not yet been accepted to a journal for publication.

***p*-value.** A statistic that gives the calculated probability that the null hypothesis could be true even given the observed differences between conditions.

randomized controlled trial. An experimental design that randomly assigns participants to the experimental or control group to better control variables across the groups.

replication. Herein, a study carried out with the intent of mimicking another to the closest degree possible.

replication crisis. A series of observations, beginning around 2010, suggesting that the results of many published scientific studies are not replicable and their conclusions potentially incorrect.

science. A body of knowledge regarding the natural world and the process for building that knowledge based on evidence acquired through observation, experiment, and simulation.

systematic review. A publication type that attempts to synthesize and summarize the current state of research on a topic.

theory. Within science, a concise, coherent, and predictive explanation for a broad range of natural phenomena that integrates and makes sense of many hypotheses, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

uncertainty. In reference to statistics, a parameter that indicates the dispersion of values within which the true value is likely to fall.

Reference Guide on Forensic Feature Comparison Evidence

VALENA E. BEETY, JANE CAMPBELL MORIARTY, AND ANDREA L. ROTH

Valena E. Beety, Robert H. McKinney Professor of Law, Indiana University Maurer School of Law.

Jane Campbell Moriarty, Carol Los Mansmann Chair in Faculty Scholarship and Professor of Law, Thomas R. Kline School of Law at Duquesne University.

Andrea L. Roth, Professor of Law and Barry Tarlow Chancellor's Chair in Criminal Justice, UC Berkeley School of Law.

CONTENTS

Overview

This reference guide discusses techniques used for forensic feature comparison, meaning comparing features in trace evidence (such as a latent fingerprint, handwriting on a document, or toolmarks on a projectile) to a reference sample (such as a suspect's reference fingerprints or handwriting exemplar, or a projectile fired from a known weapon) to determine whether the samples might originate from the same source. These techniques play a significant role in modern federal litigation, particularly criminal cases. In turn, federal judges play a central gatekeeping and managerial role in regulating the admissibility and use at trial of these techniques. The role is inherently challenging, as judges are asked to weigh in on complex issues of scientific validity that are the realm of experts. The role is also particularly challenging now, as some techniques' ability to accurately determine whether two patterns have a common source have been challenged as unreliable in light of high-profile DNA exonerations and recent governmental reports critical of feature comparison disciplines. Meanwhile, the increasing automation of feature comparison techniques with the help of software raises unique issues.

With these realities in mind, the goal of this reference guide is to provide an accessible overview for judges as they resolve legal questions about forensic feature comparison evidence. While other guides in this manual cover in depth the legal rules related to expert testimony and the specific issues related to forensic DNA typing, this reference guide offers judges a solid background in scientific and legal issues related to a number of forensic feature comparison techniques. This guide's scope is limited to feature comparison techniques, as such techniques raise recurring issues worthy of separate treatment and do not include other technical or scientific fields such as medicolegal death investigations, seized drug analysis, or digital forensics.

The first section of this reference guide explains the basic legal framework for determining the admissibility of forensic feature comparison evidence, situating the case law and rules discussed in *The Admissibility of Expert Testimony*, in this manual, in the specific context of this type of evidence.

The second section catalogs the most frequently cited institutional efforts to evaluate forensic feature comparison evidence, such as the 2009 National Research Council report (2009 NRC Report), the 2016 report of the President's Council of Advisors on Science and Technology (PCAST), the National Commission on Forensic Science (NCFS), and the National Institute of Standards and Technology (NIST)'s Organization of Scientific Area Committees (OSAC). We catalog the major findings of these groups not only because they are frequently cited by litigants but because many lower-court rulings on the admissibility of forensic feature comparison evidence were issued before these reports and may not be consistent with them.

The third section of this reference guide explains the continuing concerns that feature comparison evidence poses, with respect to experts' arguable

overclaims of certainty or individualization, frequently confused terms such as reliability versus validity, class versus individual characteristics, and other matters. This section also provides an in-depth discussion of foundational reliability and reliability as applied.

The fourth section offers a detailed explanation of some of the forensic feature comparison techniques currently in use and most likely to arise in federal cases in the next decade. These techniques include fingerprint analysis, handwriting comparison, firearms and toolmark identification, and bitemark evidence. While some techniques like bitemark analysis and handwriting comparison appear to be gradually receding in use, we include them both because they are still offered as proof and because they are sure to arise in habeas litigation or other postconviction proceedings for years to come. This guide omits several other techniques, such as shoeprint comparisons, hair and fiber analysis, and glass comparisons, that present some of the same issues as those techniques that are covered.[1] Each overview explains the basic method, the extent of empirical studies on the method, and the method's legal status in the courts. We end with a brief discussion of facial recognition and other emerging techniques.

Finally, this reference guide offers a section on automated, machine-generated feature comparison conclusions. While the *Reference Guide on Human DNA Identification Evidence*, in this manual, discusses DNA software, other feature comparison techniques are also being increasingly automated. We offer an overview of the types of machine-generated forensic conclusions a federal judge is likely to see, as well as the legal issues likely to arise in admitting these conclusions. The goal of this final section is not to definitively resolve these issues, but to equip judges to understand the stakes and assumptions underlying the conflicts they themselves will need to resolve.

Legal Framework Governing Forensic Feature Comparison Evidence

This section builds on *The Admissibility of Expert Testimony*, in this manual, by briefly highlighting significant aspects of Federal Rule of Evidence 702's application to forensic feature comparison evidence in particular. As *The Admissibility of Expert Testimony* discusses in depth, Rule 702 codifies the so-called "*Daubert* trilogy" and requires that an expert be qualified, that the opinion be helpful to the jury, that the testimony be based on sufficient facts or data, and that the testimony is the product of a method that is both foundationally reliable and has been reliably

1. Judges can find a more comprehensive list of feature comparison disciplines by perusing OSAC's website: https://perma.cc/8W8P-DHFT. Clicking on the link of each subcommittee leads to a list of proposed standards on a host of subdisciplines.

applied to the facts of the case.[2] As *Daubert* noted, courts determining the reliability of scientific expert testimony look to various factors, such as whether a method has been tested; the method's error rate; whether the method has been subject to peer review; whether the method is governed by standards and protocols; and whether the method is generally accepted in the relevant scientific community.[3]

As noted in *The Admissibility of Expert Testimony*, the 2023 amendments to Rule 702 and the advisory committee's note make explicit that the proponent must meet the requirements above by a preponderance of the evidence.[4] The advisory committee's note urges special caution with "subjective" forensic methods:

> The amendment is especially pertinent to the testimony of forensic experts in both criminal and civil cases. Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error. In deciding whether to admit forensic expert testimony, the judge should (where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results. Expert opinion testimony regarding the weight of feature comparison evidence (i.e., evidence that a set of features corresponds between two examined items) must be limited to those inferences that can reasonably be drawn from a reliable application of the principles and methods.[5]

What follows is a brief overview of each admissibility requirement in the context of feature comparison evidence.

Helpfulness to the jury. First, the skill of non-DNA feature comparison experts may well be "helpful to the jury" in explaining the evidence. For example, in a firearm comparison case, one federal district court recently stated that the expert's "technical knowledge, skill, and training to make microscopic observations of and comparisons between cartridge cases, would be helpful to the trier of fact."[6] Courts have come to similar conclusions with respect to other feature comparison disciplines.[7] But these rulings presuppose that the conclusion itself is reliable and thus worthy of being heard by the jury. In short, "helpfulness to the jury" is typically not the dispositive factor affecting admissibility of feature

2. Fed. R. Evid. 702.

3. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993) (listing nonexhaustive factors in determining reliability of a scientific expert method).

4. *See* Fed. R. Evid. 702 & advisory committee's note to 2023 amendment.

5. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

6. United States v. Davis, No. 4:18-CR-00011, 2019 WL 4306971, at *6 (W.D. Va. Sept. 11, 2019).

7. *See, e.g.*, United States v. Dale, 618 F. App'x 494 (11th Cir. 2015) (fingerprint and handwriting comparison); United States v. Shipp, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (firearm comparison); United States v. Mallory, 902 F.3d 584, 593 (6th Cir. 2018) (handwriting comparison).

comparison discipline evidence once it is deemed foundationally reliable and reliable as applied.

Expert qualifications. When allowing a feature comparison discipline expert to testify "before the jury cloaked with the mantle of an expert," a trial court must "assure that a proffered witness truly qualifies as an expert."[8] Trial courts should be careful not to conflate expert qualification with either foundational reliability or reliability as applied. A method might be foundationally reliable, but wielded by an examiner who is not sufficiently proficient to be qualified as an expert to begin with. In turn, as explained more fully below, even an expert with established credentials and proficiency must demonstrate through documentation that the method was reliably applied in the case at hand.

In theory, Rule 702 allows much flexibility to the proponent in how to show an expert is qualified. The rule is written in the disjunctive—qualified by "knowledge, skill, experience, training, or education"—and provides multiple avenues for qualifying witnesses, including experience alone.[9] Moreover, experts need not be "blue-ribbon practitioners" with "optimal qualification[s],"[10] or even "highly qualified in order to testify about a given issue."[11]

Still, courts must provide sufficient reasons for finding the expert qualified so that an appellate court can determine "whether the district court properly applied the relevant law."[12] The question is not merely whether the witness is qualified in a vacuum to speak on expert matters; it is whether the witness is qualified to offer the opinion she is offering. Some courts have excluded feature-comparison experts on this ground.[13] An expert might also be qualified to render opinions on some topics but not others.[14] Specifically, some (but not all) courts have deemed

8. United States v. Cloud, No. 1:19-CR-02032-SMJ-1, 2022 WL 801694, at *1 (E.D. Wash. Mar. 8, 2022) (fingerprint expert) (quoting Jinro Am. Inc. v. Secure Invs., Inc., 266 F.3d 993, 1004 (9th Cir. 2001)).

9. The advisory committee's note to the 2000 amendment of Rule 702 states that "[n]othing in this amendment is intended to suggest that experience alone . . . may not provide a sufficient foundation for expert testimony." Fed. R. Evid. 702 advisory committee's note to 2000 amendment.

10. United States v. Vargas, 471 F.3d 255, 262 (1st Cir. 2006).

11. Huss v. Gayden, 571 F.3d 442, 452 (5th Cir. 2009).

12. United States v. Avitia-Guillen, 680 F.3d 1253, 1259 (10th Cir. 2012) (no abuse of discretion ruling that fingerprint expert was qualified; district court provided sufficient information to establish basis for the opinion).

13. *See, e.g.*, Balimunkwe v. Bank of Am., No. 1:14-CV-327, 2015 WL 5167632 (S.D. Ohio Sept. 3, 2015) (expert not permitted to testify where he did not possess sufficient qualifications as handwriting comparison expert); Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401, 424 (S.D.N.Y 2016) (same).

14. *See, e.g.*, Marten Transp., Ltd. v. Plattform Advertising, Inc., 184 F. Supp. 3d 1006 (D. Kan. 2016) (holding that expert witness, though qualified to render several opinions on nature of trucking industry and hiring practices, was not qualified to render opinion on search engine optimization).

the exclusion of experts such as law professors, who themselves are not scientists but who have written extensively about a discipline, not an abuse of discretion.[15]

Before the 2023 amendments to Rule 702, some courts had come close to deeming a forensic feature comparison expert unqualified, but concluded that qualification was a low enough bar that any questions should go to weight, not admissibility.[16] The 2023 amendments explicitly note that, while questions of precise weight as to the opinions of a qualified expert are for the jury, the issue of qualification (as well as helpfulness and reliability) goes to admissibility and not merely weight.[17] Ultimately, appellate courts reviewing qualifications of forensic feature or pattern comparison experts have affirmed that it is the judge's obligation, and not a jury question, to determine whether a witness possesses sufficient qualifications.[18]

Some recent commentary argues that, given the 2023 amendments' concern over "subjective" methods based on examiner judgment and experience, trial courts should take proficiency testing results (or lack thereof) of feature comparison experts more seriously at the expert qualification stage.[19] Several consensus reports and other scholars have similarly urged courts to be more careful gatekeepers about the qualifications of feature comparison experts, noting that the proficiency tests that exist to date are not encouraging and the results may be misleading, as the tests do not simulate realistic crime-scene samples.[20]

15. *See, e.g.*, State v. Clifford, 121 P.3d 489 (Mont. 2005) (holding that trial court did not abuse its discretion in excluding law professor as expert in handwriting); United States v. Paul, 175 F.3d 906, 912 (11th Cir. 1999) (same).

16. *See, e.g.*, Banuchi v. City of Homestead, 606 F. Supp. 3d 1262, 1272–73 (S.D. Fla. 2022) (finding an expert "barely" qualified to testify about why fingerprints would not be at a crime scene, reasoning that "[t]he qualification standard for expert testimony is not stringent, and so long as the expert is *minimally* qualified, objections to the level of the expert's expertise [go] to credibility and weight, not admissibility") (quoting Vision I Homeowners Ass'n, Inc. v. Aspen Specialty Ins. Co., 674 F. Supp. 2d 1321, 1325 (S.D. Fla. 2009) (emphasis added) (quoting Kilpatrick v. Breg, Inc., Case No. 08-10052-CIV, 2009 WL 2058384, at *3 (S.D. Fla. June 25, 2009))); Thomas v. United States, No. 2:03-CV-02416-JPM, 2015 WL 5076969, at *184 (W.D. Tenn. Aug. 27, 2015), *aff'd*, 849 F.3d 669 (6th Cir. 2017) (explaining the witness's lack of qualifications in the area of handwriting comparison, but admitting it with the caveat that it was entitled to "little weight" in a bench trial).

17. *See* Fed. R. Evid. 702 & advisory committee's note to 2023 amendment.

18. *See, e.g.*, United States v. Ruvalcaba-Garcia, 923 F.3d 1183, 1189 (9th Cir. 2019) (trial judge erred in ruling that fingerprint technician's qualification "as an expert" was a "determination for the jury").

19. *See, e.g.*, Brandon L. Garrett & Gregory Mitchell, *The Proficiency of Experts*, 166 U. Pa. L. Rev. 901, 902 (2018) (explaining their evaluation of twenty years of fingerprint proficiency tests reflected a "surprisingly high rates of false positive identifications" and that "credentials and experience are often poor proxies for proficiency").

20. *See, e.g.*, President's Council of Advisors on Sci. & Tech. (PCAST), Exec. Office of the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods 51 (Sept. 2016), https://perma.cc/GJ9A-2DFS [hereinafter 2016 PCAST Report]; Garrett & Mitchell, *supra* note 19, at 902 (explaining their evaluation of twenty years of fingerprint proficiency tests reflected a "surprisingly high rates of false positive identifications");

Foundational reliability. This section provides a brief reminder of the foundational reliability requirements of Rule 702. A more in-depth discussion of specific recurring reliability issues in feature comparison evidence is set forth in the next section.

In determining foundational reliability of feature comparison expert testimony, judges have multiple options under Rule 702's framework, consistent with their role as gatekeeper: (1) they may decide that the evidence is admissible; (2) they may exclude the evidence entirely, concluding that the proponent has not established foundational reliability; or (3) they may admit the evidence with limitations, as some courts have done.[21] While trial court determinations of admissibility are reviewed for an "abuse of discretion,"[22] a misapprehension as to the applicable legal standard under Rule 702 is considered an abuse of discretion.[23]

One issue facing judges when applying Rule 702's reliability requirements to feature comparison disciplines is whether such disciplines are "scientific," as compared to "technical" or "specialized." On the one hand, feature and pattern recognition experts distinguish themselves from lay persons on their ability to identify individuals based on the experts' knowledge, training, and expertise. Although lay people may be able to identify a known individual's handwriting[24] or may be able to distinguish between two clearly dissimilar fingerprints, courts have consistently ruled that feature and pattern recognition by trained individuals is, in fact, specialized knowledge.[25] On the other hand, some commentators have claimed that examiners wielding relatively more subjective experience- and judgment-based

Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 Ariz. St. L. J. 1369 (2017) (explaining that, in most areas of forensic science, the necessary proficiency testing has not been done). Qualifications, proficiency, and reliability as applied are not entirely distinct concepts, and there is some overlap among them.

21. *See, e.g.*, United States v. Tibbs, No. 2016-CF1-19431, 2019 WL 4359486 (D.C. Super. Ct. Sept. 5, 2019) (precluded the government from eliciting testimony identifying the recovered firearm as the source of the recovered cartridge and limiting expert's testimony to a conclusion that the firearm cannot be excluded as the source, based on the consistency of the class characteristics and microscopic toolmarks); United States v. Adams, 444 F. Supp. 3d 1248 (D. Or. 2020); United States v. Shipp, 422 F. Supp. 3d 762, 778–79 (E.D.N.Y. 2019) (disallowing a conclusion of a "match" with firearm evidence); United States v. Rutherford, 104 F. Supp. 2d 1190, 1193 (D. Neb. 2000) (disallowing the conclusion of a "match" of handwriting samples); United States v. Van Wyk, 83 F. Supp. 2d 515 (D.N.J. 2000); United States v. Hines, 55 F. Supp. 2d 62 (D. Mass. 1999) (same).

22. *See* Gen. Elec. Co. v. Joiner, 522 U.S. 136, 141–43 (1997).

23. *See, e.g.*, Koon v. United States, 518 U.S. 81, 100 (1996) (noting that a trial court necessarily abuses its discretion when it fails to apply the correct legal standard or makes an error of law); Jeff D. v. Otter, 643 F.3d 278 (9th Cir. 2011) (same).

24. *See, e.g.*, Fed. R. Evid. 901(b)(2) (permitting "nonexpert's opinion that handwriting is genuine, based on a familiarity with it that was not acquired for the current litigation").

25. *See, e.g.*, *Shipp*, 422 F. Supp. 3d at 783 (firearm comparison); United States v. Llera Plaza, 188 F. Supp. 2d 549, 563–64 (E.D. Pa. 2002) (fingerprint comparison); United States v. Mallory, 902 F.3d 584, 593 (6th Cir. 2018) (handwriting comparison).

feature and pattern identification techniques are not "scientists" practicing "science."[26]

Of course, a field that is "technical" rather than "scientific" is still subject to the reliability requirements of Rule 702, as the Supreme Court held in *Kumho Tire Co. v. Carmichael*.[27] But the Supreme Court has not squarely addressed the types of factors a trial judge should consider in determining the foundational reliability of a nonscientific technical method.[28] The National District Attorneys Association suggested in its response to the 2016 PCAST Report that validation studies are not required for nonscientific, experience-based methods.[29] But as the PCAST committee responded, it is not clear what the alternative means of testing validity of such a method would be.[30] While some have argued that validity can be established through so-called "blind" verification by a second examiner or interlaboratory comparison, such comparisons would presumably focus mostly on repeatability, not accuracy. Thus, courts determining the foundational reliability of a feature comparison discipline method deemed nonscientific will have to determine how—if not through validation studies showing the limits of the method's ability to get the right answer—the proponent of the method can establish its accuracy for its stated purpose by a preponderance of the evidence.

Reliability as applied. Once the proponent has demonstrated that the method is foundationally reliable (i.e., the method produces accurate, repeatable, and reproducible results on materials like those in the instant case), and that the examiner is qualified to conduct the method, the proponent must introduce additional evidence that the examiner in fact followed the method in the case at hand. The 2023 amendments to Rule 702 reaffirmed that reliability as applied, like other admissibility requirements, goes to admissibility and not merely weight and must be proven by a preponderance of the evidence.[31] To prove that

26. *See* Letter from Nat'l Dist. Att'ys Ass'n to President Obama in response to 2016 PCAST Report (Nov. 16, 2016), https://perma.cc/5JX7-BU5V. *See generally* discussion in section titled "Recurring Concerns with Forensic Feature Comparison Evidence" below.

27. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 142 (1999).

28. *See generally* discussion in section titled "Prior Government and NGO Review of Forensic Feature Comparison Evidence" below (discussing critiques of PCAST and arguments about how and whether to calculate error rates for feature comparison disciplines).

29. *See id.* (noting the NDAA's response to the 2016 PCAST Report).

30. *See* President's Council of Advisors on Sci. & Tech. (PCAST), Exec. Office of the President, An Addendum to the PCAST Report on Forensic Science in Criminal Courts 1, 3 (Jan. 6, 2017), https://perma.cc/TFP8-YUYS [hereinafter PCAST Addendum] (Noting that some critics of PCAST "suggested that the validity and reliability of such a method could be established without actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any alternative approach that could establish the validity and reliability of a subjective forensic feature-comparison method.").

31. Fed. R. Evid. 702 advisory committee's note to 2023 amendment ("[M]any courts have held that the critical questions of the . . . application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a).").

a laboratory or examiner reliably applied a method, the proponent might rely on various types of documentation, from internal validation studies (showing that the method works on the laboratory's equipment with the specific protocols, practices, and personnel of the specific forensic service provider) to case file documentation to examiner bench notes, reports, and testimony showing the conditions of the case and what the examiner did in the particular case.

This "reliability as applied" showing is different from foundational reliability or expert qualification. For example, a trial court might deem fingerprint comparison a foundationally reliable method and deem a particular examiner to be qualified based on experience, credentials, and proficiency tests. But if the comparison in the case is between a reference print and a particularly smudged or incomplete print that goes beyond the bounds of the method's empirically tested accuracy, the method might not be reliably applied. As the advisory committee note to the 2023 amendments emphasized, expert opinions must "stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology."[32] Another way that an opinion can exceed the bounds of a valid application of a method is if it relates to a different type of conclusion—for example, whether a note was made to look like a forgery, rather than whether two notes have the same author.[33] Still another way an opinion might be unreliable as applied is if the examiner claims certainty in their opinion, a concern the advisory committee considers "particularly pertinent to the testimony of forensic experts." Instead, in deciding whether to admit the evidence, trial courts should "(where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results."[34]

Prior Government and NGO Review of Forensic Feature Comparison Evidence

This subsection offers an overview of governmental and nongovernmental organization (NGO) review of forensic feature comparison evidence over the past fifteen years. Judges should be aware of these reports both because they are frequently cited and because they can be helpful guides to areas of continuing research and controversy.

32. *See id.*

33. *See, e.g.*, Almeciga v. Ctr. for Investigative Reporting, Inc, 185 F. Supp. 3d 401, 424 (S.D.N.Y 2016) (noting that method for one purpose was not validated for other purpose and that a trial judge "should not [admit questioned document testimony] without carefully evaluating whether the examiner has actual expertise in regard to the specific task at hand").

34. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

The 2009 National Research Council Report. The push since the mid-2000s to improve feature comparison disciplines has resulted in part from the DNA exoneration movement, which exposed erroneous convictions based on embellished or misleading forensic evidence.[35] Additionally, the misidentification of Brandon Mayfield by the FBI as the perpetrator of the 2004 Madrid train bombings based on latent fingerprint analysis further spurred concern over non-DNA methods.[36] The National Research Council of the National Academies (NRC) began a formal review in 2006 of the state of non-DNA forensic feature comparison disciplines.[37] The result was a lengthy report published in 2009, *Strengthening Forensic Science in the United States: A Path Forward*.[38]

The primary conclusion of the 2009 NRC Report was that, "[w]ith the exception of nuclear DNA analysis . . . no forensic [feature identification] method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source."[39] Specifically, the NRC pointed out the relative subjectivity and nonreproducibility in non-DNA methods that are based primarily on examiner experience and judgment; the lack of validation studies to assess accuracy of these relatively subjective methods;[40] the lack of "population studies" that show the "variability" and rarity of a feature within a relevant population (critical to assessing the probative value of a "match" between two sets of observed features, like whorls and loops in a thumbprint or a certain shoe sole pattern);[41] the lack of corrective measures to avoid examiner contextual bias;[42] the "absence of a feedback mechanism" to alert examiners that a method

35. *See generally* Restoring Freedom, Innocence Project, https://perma.cc/LMU4-87JU (last visited Nov. 18, 2024) (explaining the categories of reasons underlying convictions of factually innocent defendants).

36. *See* Robert B. Stacey, Report on the Erroneous Fingerprint Individualization in the Madrid Train Bombing Case, Fed. Bureau of Investigation (FBI) (Jan. 2005), https://perma.cc/JQ9W-RX7K.

37. *See generally* Erin Murphy, *What 'Strengthening Forensic Science' Today Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report*, 9 L. Probability & Risk 7 (2010), https://doi.org/10.1093/lpr/mgp030 (explaining the story of how DNA exonerations led to the Senate's charge, and the forensic community's request, to the NRC to write a report on non-DNA evidence); Nat'l Rsch. Council, Nat'l Acads., Strengthening Forensic Science in the United States: A Path Forward 1 (2009), https://doi.org/10.17226/12589 [hereinafter 2009 NRC Report] (explaining Congress's charge to them). We refer to the 2009 report as the "2009 NRC Report" to distinguish it from later reports from the National Academy of Sciences, even though some authors refer to the 2009 report as the "NAS Report."

38. *See* 2009 NRC Report, *supra* note 37.

39. *Id*. at 7.

40. *Id*. at 8.

41. *Id*. at 149. *See also id*. at 154 (same concern for firearms and toolmarks), 163 (same for fiber evidence), 165 (handwriting), 184 (same concern for friction ridge patterns).

42. *Id*. at 8, 24.

might have produced a false positive or negative;[43] and the lack of "quantifiable measures of uncertainty" (i.e., an error rate).[44]

Responses to the 2009 NRC Report varied, although the scientific community appeared largely supportive. Some pushed back against the report's calling established methods into question.[45] On the other hand, organizations such as the American Academy of Forensic Sciences (AAFS) published a formal response "support[ing] the recommendations" of the report,[46] and other scientists followed suit.[47]

2013–17: The National Commission on Forensic Science (NCFS). The 2009 NRC Report led to a broad debate over the validity and reliability of forensic science methods, and to the creation of the NCFS, a joint project of the Department of Justice (DOJ) and NIST. The NCFS's thirty-seven members included forensic examiners, psychologists and other scientists, prosecutors, defense attorneys, academics, and judges.[48] Between 2014 and 2017, the NCFS approved twenty "recommendations" and twenty-three "views documents," all available on the Commission's website[49] and including recommendations requiring proficiency testing of examiners even in fields lacking an accreditation body; views on what federal prosecutors should disclose in pretrial discovery about forensic methods, above and beyond what is required by Federal Rule of Criminal Procedure 16;[50] and the view that forensic examiners should base their opinions only on

43. *Id.* at 149.

44. *Id.* at 23.

45. *See generally* Paul C. Giannelli, *The 2009 NAS Forensic Science Report: A Literature Review*, 48 Crim. L. Bulletin 378, 382, 385 (2012) (discussing responses to the 2009 NRC Report, including a prosecutor in front of the Senate Judiciary Committee insisting that the report was an "agenda-driven attack upon well-founded investigative techniques" and then-Senator Jeff Sessions' comment that "I don't think we should suggest that those proven scientific principles that we've been using for decades are somehow uncertain").

46. *See* Am. Acad. of Forensic Scis. (AAFS), Response to the National Academy of Sciences' "Forensic Needs" Report (Sept. 4, 2009), https://perma.cc/K7B4-TWCP.

47. *See, e.g.*, Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 Int'l Statistical Rev. 111, 120 (2015), https://doi.org/10.1111/insr.12069 (Expressing support for the 2009 NRC Report and opining that "[t]he scientific method is very general, and the principles of science apply to all branches, irrespective of the specific application (e.g. physics, chemistry or biology). If the scientific method has not been applied, then scientists from no branch can defend the results.").

48. *See* U.S. Dep't of Justice, National Commission on Forensic Science: Members (June 1, 2017), https://perma.cc/V98X-BMQB (listing members and biographies).

49. *See* U.S. Dep't of Justice, National Commission on Forensic Science: Work Products Adopted by the Commission (July 23, 2018), https://perma.cc/5XVZ-KN4F (links to documents).

50. *See id.* These recommendations were written before the 2022 amendments to Rule 16, which (among other things) require parties to disclose not only written summaries of their expert's anticipated testimony but "a complete statement" of the witness's opinions, the bases and reasons for those opinions, the witness's qualifications, and a list of other cases in which the witness had testified in the previous four years. Fed. R. Crim. P. 16 advisory committee's note to 2022 amendment.

"task-relevant" information.[51] While the NCFS's charter expired in April 2017 and was not renewed by then-Attorney General Jeff Sessions,[52] the views documents remain potentially persuasive documents wielded by parties litigating forensic evidence issues.

The 2016 PCAST Report. While the NCFS was still completing its work, PCAST issued a report in 2016 on forensic feature comparison evidence.[53] The report's working group largely consisted of judges and academics from the "derivative sciences" (i.e., sciences like chemistry, statistics, genetics, and computer science underlying forensic methods), although the group sought input from "the Federal Bureau of Investigation (FBI) Laboratory and individual scientists at NIST, as well as from many other forensic scientists and practitioners, judges, prosecutors, defense attorneys, academic researchers, criminal-justice-reform advocates, and representatives of Federal agencies."[54]

The 2016 PCAST Report echoed the concerns of the 2009 NRC Report and focused on explaining what was missing to properly validate non-DNA forensic feature comparison methods. Specifically, PCAST explained that the only way to meaningfully test the validity of a method based primarily on examiner judgment and experience (where the method's steps are largely an inscrutable "black box" to those on the outside) is through "black box" studies where the method's accuracy is tested on samples where the tester knows the right answer.[55] In turn, PCAST explained that any statement about a method's reliability, or the certainty underlying an examiner's conclusion, should not overstate the probative value of the evidence beyond what is "demonstrated by empirical evidence" based on error-rate studies.[56] "For example," PCAST wrote, "if the false positive rate of a method has been found to be 1 in 50, experts should not imply that the method is able to produce results at a higher accuracy."[57]

The PCAST Report then covered what a well-designed error rate study requires: samples and populations representative of real casework; a sufficiently

51. *See* U.S. Dep't of Justice, Views of the Commission Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information (Dec. 8, 2015), https://perma.cc/U2JF-QSCP. This document was heavily influenced by, and frequently cited, the work of psychologist Itiel Dror (who spoke to the commission), who has written on contextual bias in forensic feature comparison methods such as DNA mixture interpretation and latent fingerprint analysis. *See* Itiel E. Dror & Greg Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, 51 Sci. & Justice 204 (2011), https://doi.org/10.1016/j.scijus.2011.08.004, *and* Itiel E. Dror et al., *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a 'Target' Comparison*, 208 Forensic Sci. Int'l 10 (2011), https://doi.org/10.1016/j.forsciint.2010.10.013.

52. *See* U.S. Dep't of Justice, National Commission on Forensic Science, https://perma.cc/V6GT-YSWN (noting charter expiration in April 2017).

53. *See* 2016 PCAST Report, *supra* note 20.

54. *Id.* at 2.

55. *Id.* at 5.

56. *Id.* at 6.

57. *Id.*

large sample size; "blind" testing; and no mid-stream change in protocol. In turn, PCAST opined on the proper way to calculate a method's error rate. First, it insisted that the false positive rate be calculated by determining the number of false inclusions among an examiner's *conclusive* examinations (exclusions or declarations of a "match") rather than all examinations (including "inconclusive" determinations).[58] Second, the reported error rate should be the upper bound of a "confidence interval" around the study error rate to account for measurement uncertainty, so that the factfinder has a sense of how high the false positive rate might be.[59] Using these criteria, PCAST ultimately found insufficient evidence of foundational validity for several feature comparison disciplines—toolmark analysis, handwriting analysis, bitemark analysis, footwear analysis, microscopic hair analysis, and deconvolution of DNA mixtures over three contributors.[60] The only non-DNA feature comparison method PCAST deemed foundationally valid was latent print analysis, but added several caveats.[61]

The 2016 PCAST Report garnered formal responses from numerous organizations, some supportive and some critical.[62] PCAST issued an addendum to its report in January 2017 responding to critiques.[63] After the addendum, additional critiques[64] and defenses[65] of PCAST continued. A 2018 article by Professor Jay Koehler, directed at the judiciary, offers a helpful and measured summary of the critiques of PCAST, possible responses, lingering debates, and issues for

58. *Id.* at 51.

59. *Id.*

60. *Id.* at 7.

61. *Id.* at 9–10.

62. *See, e.g.*, Nat'l Dist. Att'ys Ass'n, *supra* note 26 (arguing that not all methods are "scientific," and not all methods need to be supported by validation studies); Fed. Bureau of Investigation (FBI), Comments on President's Council of Advisors on Science and Technology REPORT TO THE PRESIDENT Forensic Science in Federal Criminal Courts: Ensuring Scientific Validity of Pattern Comparison Methods [PCAST Report] (Sept. 20, 2016), https://perma.cc/B5S9-XSG3 (criticizing PCAST for not taking into account studies that the FBI argued met the criteria for appropriate black box studies).

63. *See* PCAST Addendum, *supra* note 30.

64. *See, e.g.*, Ted Robert Hunt, *Scientific Validity and Error Rates: A Short Response to the PCAST Report,* 86 Fordham L. Rev. Online 24, 26 (2018) (agreeing that methods must be tested empirically but deeming PCAST's views of an appropriate test as too narrow); U.S. Dep't of Justice, *Justice Department Publishes Statement on 2016 President's Council of Advisors on Science and Technology Report* (Jan. 13, 2021), https://perma.cc/8J6L-L4DN (arguing that black box studies are not always needed and that PCAST's criteria are too narrow).

65. *See* Letter from Democracy Forward Foundation on behalf of Union of Concerned Scientists, requesting correction under the Information Quality Act (regarding U.S. Dep't of Justice's statement on the PCAST Report) (June 24, 2021), https://perma.cc/MQD8-K835 (defending PCAST and demanding that DOJ retract its statement); Innocence Project Staff, *Innocence Project Calls on Department of Justice to Retract Statement on PCAST Report* (Feb. 19, 2021), https://perma.cc/SS23-4K5F (same).

judges to resolve in light of these debates.[66] For their part, statisticians and other scientists have likewise urged the forensic community to follow PCAST's recommendations.[67]

Notwithstanding its continued criticism of PCAST, DOJ has voluntarily made several efforts to improve its forensic practices in response to PCAST and other reports. From 2017 onward, the department has implemented several new initiatives, including an updated code of professional responsibility, publication of research needs, testimony monitoring, and creation of a laboratory needs working group.[68] FBI representatives have also reported such DOJ initiatives being implemented to strengthen the accuracy of FBI experts' testimony.[69] DOJ does continue to use and endorse the term "source identification," notwithstanding concerns expressed by others to the Advisory Committee on the Rules of Evidence that this term is synonymous with a claim of individualization, which DOJ reports its experts no longer testify to in court.[70]

The forensic community has responded to PCAST in part by promulgating standards to improve feature comparison disciplines through a new entity, NIST's Organization of Scientific Area Committees (OSAC). OSAC is composed of hundreds of forensic examiners along with lawyers, judges, psychologists, statisticians, and laboratory directors. Created in 2014, its mission is "to strengthen the nation's use of forensic science by facilitating the development

66. Jonathan Jay Koehler, *How Trial Judges Should Think About Forensic Science Evidence*, 102 Judicature 29 (2018), https://perma.cc/DLM4-CXCL.

67. *See, e.g.*, Suzanne Bell et al., *A Call for More Science in Forensic Science*, 115 Proc. Nat'l Acad. Scis. 4541, 4541 (2018), https://doi.org/10.1073/pnas.1712161115 (Discussing the 2009 NRC and PCAST reports with approval and opining that "[f]orensic science is at a crossroads. It is torn between the practices of science, which require empirical demonstration of the validity and accuracy of methods, and the practices of law, which accept methods based on historical precedent even if they have never been subjected to meaningful empirical validation."); Thomas D. Albright, *The US Department of Justice Stumbles on Visual Perception*, 118 Proc. Nat'l Acad. Scis. e2102702118 (2021), https://doi.org/10.1073/pnas.2102702118 (offering a neuroscientist's critique of DOJ's refusal to adopt PCAST's findings).

68. *See* U.S. Dep't of Justice, Forensic Science: Publications, https://perma.cc/BLR3-WNTK (last visited Nov. 20, 2024) (listing priorities, publications, and initiatives).

69. *See, e.g.*, Alice R. Isenberg & Cary T. Oien, *Scientific Excellence in the Forensic Science Community*, 86 Fordham L. Rev. Online 39 (2018) (discussing testimony monitoring as an example of a recent initiative, as well as the FBI's existing accreditation requirements); Andrew D. Goldsmith, *The Reliability of the Adversarial System to Assess the Scientific Validity of Forensic Evidence*, 86 Fordham L. Rev. Online 16 (2018) (noting that the FBI has eliminated use of the terms "reasonable degree of scientific certainty" and similar statements; eliminated claims of zero error; prohibited examiners from citing the number of examinations conducted as an indication of the accuracy of their conclusion; and created a testimony monitoring program). However, DOJ's policies apply only to its own witnesses. If witnesses from local or state agencies testify in a federal case, they may not adhere to the DOJ policy.

70. *See* Judicial Conference Advisory Committee on the Federal Rules of Evidence, Minutes of the Meeting of May 3, 2019, 1, 20–23 (May 3, 2019), https://perma.cc/2QRN-7LQ8 (disagreement on the FBI's continued use of this term and whether it is different from individualization).

and promoting the use of high-quality, technically sound standards."[71] OSAC both creates new standards and reviews standards set by other standards development organizations (SDOs), such as the American Academy of Forensic Science's (AAFS) Academy Standards Board (ASB),[72] ASTM International (formerly the American Society of Testing and Materials),[73] and the National Fire Protection Association (NFPA).[74] Standards that have sufficient "technical merit" are placed on OSAC's "registry."[75] Judges should be aware of OSAC's activities and output, as well as controversy over the quality of its registry standards, given that a standard's inclusion on the registry may be touted by one side or another as proof that a discipline is governed by standards, one of the "*Daubert* factors."[76]

In addition to OSAC, other governmental and nongovernmental organizations independent of the forensic examiner or legal communities[77] continue efforts to improve forensic science. The Center for Statistics and Applications in

71. NIST's Organization of Scientific Area Committees for Forensic Science (OSAC), *About Us*, Nat'l Inst. of Standards & Tech. (NIST), U.S. Dep't of Commerce, https://perma.cc/9NJ8-7L2F (last visited Nov. 20, 2024).

72. *See* Academy Standards Board, Am. Acad. of Forensic Sci. (AAFS), https://perma.cc/U74R-Y4SM (last visited Nov. 20, 2024).

73. *See* ASTM International, https://perma.cc/25VC-ZCQK (last visited Nov. 20, 2024).

74. *See* Nat'l Fire Prot. Ass'n (NFPA), NFPA Codes and Standards, https://perma.cc/G24D-JGNZ (last visited Nov. 20, 2024).

75. *See* NIST's OSAC, *About Us, supra* note 71 ("OSAC also reviews standards and posts high quality ones to the OSAC Registry [https://perma.cc/L2SZ-L7X3 (last visited Nov. 20, 2024)]. Inclusion on this registry indicates that a standard is technically sound and that laboratories should consider adopting them.").

76. Three OSAC scientists published commentary criticizing OSAC for promulgating "vacuous" standards that merely require laboratories to *have* standards, rather than offering meaningful guidance on what the standards should be. These critics claimed that placing such standards on the OSAC registry allows disciplines to misleadingly claim they are governed by standards. Geoffrey Stewart Morrison, Cedric Neumann & Patrick Henry Geoghegan, *Vacuous Standards—Subversion of the OSAC Standards-Development Process*, 2 Forensic Sci. Int'l: Synergy 206 (2020), https://doi.org/10.1016/j.fsisyn.2020.06.005. In response, several members of ASB argued that the standards were an improvement, had to go through several rounds of comments from task groups and the public, and can be revised. *See* Linton A. Mohammed et al., *Response to Vacuous Standards Subversion of the OSAC Standards Development Process*, 3 Forensic Sci. Int'l: Synergy 100145 (2021), https://doi.org/10.1016/j.fsisyn.2021.100145. A reply to the response, signed by eleven scientists, including several OSAC members, argued that the issue is the content of the standards, not the process that led to them, and ended with an admonition to judges to "not accept at face value claims of scientific validity based on the fact that published standards have been followed. We would encourage courts to enquire further so as to ascertain whether those standards are fit for purpose." *See* Geoffrey Stewart Morrison et al., *Reply to Response to Vacuous Standards—Subversion of the OSAC Standards-Development Process*, 3 Forensic Sci. Int'l: Synergy 100149 (2021), https://doi.org/10.1016/j.fsisyn.2021.100149.

77. There are also professional organizations within the forensic examiner and legal communities that work to improve and critique forensic techniques, such as the Forensic Justice Project, https://perma.cc/RS96-DK5U, and the American Academy of Forensic Sciences, https://perma.cc/CSF3-69L4.

Forensic Evidence (CSAFE), a federally funded research and training center that is the only one of three NIST Centers of Excellence dedicated to building the scientific and statistical foundations of feature and pattern comparison evidence, remains active in nationwide efforts to review and improve forensic feature comparison methods. In particular, many of its members are statisticians developing more objective means of rendering feature and pattern comparison based on automated methods.[78] Likewise, the American Association for the Advancement of Science (AAAS) holds conferences and publishes articles related to forensic science issues.[79]

Recurring Concerns with Forensic Feature Comparison Evidence

Before this reference guide delves into specific disciplines, this section flags several recurring issues with respect to many non-DNA feature comparison methods that judges will face.

Terminology and Testimony Issues

Overclaims of certainty and challenges to claims of "matches" or source attribution. Challenges to non-DNA feature comparison expert testimony sometimes relate to the terms used by experts to describe their conclusions. One issue is claims of certainty. The advisory committee's notes to the 2023 amendments to Rule 702 state that "[f]orensic experts should avoid assertions of absolute or one hundred percent certainty."[80] The 2009 NRC Report and 2016 PCAST Report also urged against such statements,[81] and DOJ's guidelines prohibit them.

Another frequently challenged word in feature comparison reports and testimony is "match." When an expert deems two sets of features to be consistent

78. *See generally* Center for Statistics and Applications in Forensic Evidence (CSAFE), https://perma.cc/UW29-6XDB (linking to studies and ongoing research).

79. For example, the AAFS Center for Scientific Responsibility and Justice held a conference on forensics in 2019 on the tenth anniversary of the 2009 NRC Report. *See* Am. Ass'n for the Advancement of Sci. (AAAS), An Update on Strengthening Forensic Science in the United States: A Decade of Development [Forensic Conference 2019], https://perma.cc/8NXY-9WHK. *See generally* Am. Ass'n for the Advancement of Sci. (AAAS), "Forensic" Search Results, https://perma.cc/9ZRZ-4RQH (listing publications and events).

80. Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

81. *See, e.g.*, 2009 NRC Report, *supra* note 37, at 47–48 (noting testimony of some experts as to "unfailing certainty" and describing "failure to acknowledge uncertainty" as a problem in feature comparison methods); PCAST Report, *supra* note 20 at 6 (criticizing experts' "[s]tatements claiming or implying greater certainty than demonstrated by empirical evidence").

with each other, they might testify that the evidence and reference samples "match." But the term "match" might inaccurately suggest both an air of certainty in the expert's determination of the consistency features, and that the mere consistency of a set of compared features means that the two items necessarily share a common source. For example, if two fingerprints "match" at six points, they might not "match" if the examiner looked at ten more points. For a discipline like DNA, with robust population data and quantified "match" statistics, examiners do not need to fall back on terms like "match" or "identification"; they can just present a statistic and allow the factfinder to make their own determination of whether a suspect is the source of the DNA. But for more subjective methods without such data, an examiner's opinion that two patterns "match" or share the same source is more fraught.

For similar reasons, government and NGO reports have also voiced concern over use of other terms that imply that two patterns definitely have the same source, such as "individualization," "similar in all respects tested," or "to the exclusion of all others," given their "profound effect on how the trier of fact in a criminal or civil matter perceives and evaluates scientific evidence."[82] For example, the 2009 NRC Report expressed concern that imprecise terminology in microscopic hair analysis, such as "associated with," could imply individualization and a "match" without necessary confirmation from mitochondrial DNA (mtDNA) analysis.[83] Indeed, an internal FBI study found that "of 80 hair comparisons that were 'associated' through microscopic examinations, 9 of them (12.5 percent) were found in fact to come from different sources when reexamined through mtDNA analysis."[84] The 2016 PCAST Report likewise urged that "quantitative information about the reliability of methods be stated clearly in expert testimony," and ambiguous language such as "match" and "identification" be avoided, as this could be misinterpreted as scientifically supported individualization.[85] PCAST also discouraged examiners from testifying as to the "uniqueness" of features rather than focusing on the extent to which evidence of consistency in features supports an inference that the features or patterns share a common source.[86]

Other scientific organizations, such as the American Statistical Association (ASA), have voiced similar warnings about language suggesting identification or source attribution.[87] To avoid factfinders making inaccurate inferences about

82. 2009 NRC Report, *supra* note 37, at 21. *See also id.* at 7, 43 ("individualization"); 176 ("exclusion of all others").

83. *Id.* at 161.

84. *Id.* at 160–61 (discussing Max Houck & Bruce Budowle, *Correlation of Microscopic and Mitochondrial DNA Hair Comparisons*, 47 J. Forensic Scis. 964 (2002)).

85. 2016 PCAST Report, *supra* note 20, at 121–22.

86. *Id.* at 61–62.

87. *See* Am. Stat. Ass'n, American Statistical Association Position on Statistical Statements for Forensic Evidence 1, 4–5 (Jan. 2, 2019), https://perma.cc/GHH3-VY4G ("The ASA strongly discourages statements to the effect that a specific individual or object is the source of the forensic

certainty, the ASA further suggests requiring experts to make clear in their testimony that "it is possible that other individuals or objects may possess or have left a similar set of observed features" and to acknowledge "both in testimony and in written reports" a method's "absence of models and empirical evidence."[88] While some testimony guides direct examiners to limit their claims to stating that evidence is "strong support" for identification (as compared to a statement of categorical source attribution), such statements themselves should be backed up by empirical data about the rarity of features such that the examiner's observation of them justifies such language in a given case.

Still other terms have been critiqued as simply inappropriate. For example, the NCFS published a consensus views document that terms like "reasonable degree of scientific certainty" and "reasonable degree of ballistic certainty" have no scientific basis and should not be used by legal actors or examiners.[89] The NCFS also published a consensus views document on "Inconsistent Terminology," which discussed inconsistent and misleading uses across disciplines of words like "perimortem," "inconclusive," or implicitly overclaiming phrases like "there could be another individual somewhere in the world" with the same feature or pattern.[90] In federal court, some examiners will be limited by DOJ's guidelines (relatively new as of this writing) on uniform language for testimony and reports, which typically require examiners to report a source identification, a source exclusion, or an inconclusive determination.[91] Some commentators urge a retreat from such categorical language, given the lack of data about the rarity of features in the population that would presumably be needed to scientifically support any claim of individualization. Instead, some commentators urge examiners to only use scientifically supported statements about the strength of the evidence, in the form of likelihood ratios or other statements that take into account population data showing the rarity of features in the relevant population.[92]

science evidence. Instead, the ASA recommends that reports and testimony make clear that, even in circumstances involving extremely strong statistical evidence, it is possible that other individuals or objects may possess or have left a similar set of observed features. . . . The ASA recommends that the absence of models and empirical evidence be acknowledged both in testimony and in written reports.").

88. *Id*.

89. *See* U.S. Dep't of Justice, Nat'l Comm'n on Forensic Sci. (NCFS), Work Products Adopted by the Commission (July 23, 2018), https://perma.cc/64DR-4VDA (links to documents).

90. *See* U.S. Dep't of Justice, NCFS, Final Draft Views on Inconsistent Terminology, https://perma.cc/9CT6-VY3U (last visited Nov. 26, 2024).

91. *See, e.g.*, U.S. Dep't of Justice, Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Pattern Examination (2023), https://perma.cc/4N57-TDAG.

92. *See, e.g.*, Geoffrey Stewart Morrison et al., *Calculation of Likelihood Ratios for Inference of Biological Sex from Human Skeletal Remains*, 3 Forensic Sci. Int'l: Synergy 100202 (2021), https://doi.org/10.1016/j.fsisyn.2021.100202 (urging use of likelihood ratios (LRs) instead of statements of

Conflating related but distinct scientific principles such as "reliability" and "validity." Although courts often use the terms *validity* and *reliability* interchangeably, the terms have distinct meanings in different scientific disciplines. Validity typically refers to the ability of a test to "accurately measure[] what it is intended to measure."[93] Reliability, in scientific and statistical terms, refers to the extent to which a method "produces largely consistent results when properly applied."[94] Reliability can be further broken down into repeatability ("intra-examiner reliability," or the extent to which the same examiner reaches the same results under the same conditions) and reproducibility ("inter-examiner reliability," or the extent to which other examiners reach the same result using the same method).[95] The Supreme Court acknowledged these distinctions in *Daubert*, but the Court indicated that it was using the term reliability in a different sense. The Court wrote that its concern was "evidentiary reliability—that is, trustworthiness. . . . In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity.*"[96] Thus, judges should be aware that when the 2016 PCAST Report refers to "foundational validity" and "validity as applied," it is speaking to the same questions as *Daubert*. At the same time, other experts might speak of a method's reliability solely in terms of its reproducibility and repeatability, rather than its accuracy in getting the right answer. Judges should be aware of the context in which the terms are used.

"Class" versus "subclass" versus "individual" characteristics. Many non-DNA forensic feature comparison techniques purport to distinguish among "class," "subclass," and "individual" characteristics. Class and sub-class characteristics are believed to be shared by a group of persons or objects (e.g., ABO blood types, or the toolmark patterns seen on a particular kind of Glock handgun).[97] So-called individual characteristics are thought to be unique to an object or person, to the

source identification); David H. Kaye, *The Nikumaroro Bones: How Can Forensic Science Assist Fact-finders?*, 6 Va. J. Crim. L. 101 (2018) (same).

93. *See* Hal S. Stern, Maria Cuellar & David Kaye, *Reliability and Validity of Forensic Science Evidence*, 16 Significance 21, 22 (2019), https://doi.org/10.1111/j.1740-9713.2019.01250.x.

94. *Id.*

95. *Id.* at 22–23.

96. Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 590 n.9 (1993) ("We note that scientists typically distinguish between 'validity' (does the principle support what it purports to show?) and 'reliability' (does application of the principle produce consistent results?) . . . .").

97. *See* Jan S. Bashinski & Joseph L. Peterson, *Forensic Sciences*, *in* Local Government: Police Management 556 n.74 (William Geller & Darrel Stephens eds., 4th ed. 2004):

> The forensic scientist first investigates whether items possess similar 'class' characteristics—that is, whether they possess features shared by all objects or materials in a single class or category. (For firearms evidence, bullets of the same caliber, bearing rifling marks of the same number, width, and direction of twist, share class characteristics. They are consistent with being fired from the same *type* of weapon.) The forensic scientist then attempts to determine an item's 'individuality'—the features that make one thing different from all others similar to it, including those with similar class characteristics.

exclusion of all others. The term *match* is especially ambiguous with respect to these types of methods because a match could mean merely that the two patterns are thought to come from the same class or subclass, or it could mean that the two patterns are thought to come from the same source because they share individual characteristics. Expert opinions involving "individual," "subclass," and "class" characteristics raise different issues, including whether the determination rests on a firm scientific foundation[98] and how large a class or subclass is (for determining the probative value of a determination that two compared items are part of the same class or subclass).[99]

Where an examiner seeks to testify to observing an individual characteristic, judges might further inquire into the assumptions underlying that description as part of a *Daubert* inquiry. The examiner is presumably assuming that a particular observed feature is unique, raising the need for empirical data suggesting uniqueness. The examiner's use of that term also presumably reflects an assumption that a single "individual" characteristic is sufficient to make a "source identification" conclusion, given that no other item in the world has that one characteristic. Some disciplines might be able to sustain that assumption, while others might not.

Presenting conclusions in quantitative terms based solely on examinations conducted by the examiner. Forensic examiners from some disciplines and laboratories sometimes appeal to the sheer number of examinations they have done as a sufficient empirical basis for a source conclusion statement. For example, an examiner might claim that "in 10 years of practice they have examined 5,000 items like this and have never seen such a high degree of similarity," or that "the probability of observing such a strong similarity if items do not have a common source is less than 1 in a 1,000." The FBI, for its part, has recently affirmed that its testimony monitoring program seeks to eliminate such claims made by testifying experts.[100]

The fallacy of the transposed conditional (the "prosecutor's fallacy"). Forensic examiners, like all people, occasionally fall prey to statistical fallacies in drawing conclusions. One recurring example in the feature comparison context is the so-called fallacy of the transposed conditional. A conditional probability is the

98. *See* Michael Saks & Jonathan Koehler, *The Individualization Fallacy in Forensic Science Evidence*, 61 Vand. L. Rev. 199 (2008).

99. *See* Margaret A. Berger, *Procedural Paradigms for Applying the* Daubert *Test*, 78 Minn. L. Rev. 1345, 1356–57 (1994):

> We allow eyewitnesses to testify that the person fleeing the scene wore a yellow jacket and permit proof that a defendant owned a yellow jacket without establishing the background rate of yellow jackets in the community. Jurors understand, however, that others than the accused own yellow jackets. When experts testify about samples matching in every respect, the jurors may be oblivious to the probability concerns if no background rate is offered, or may be unduly prejudiced or confused if the probability of a match is confused with the probability of guilt, or if a background rate is offered that does not have an adequate scientific foundation.

100. *See* Goldsmith, *supra* note 69 (noting the FBI's position on this).

probability of one event A "given" another event B—for example, the probability that a playing card is a queen, given that it is a heart, is 1 in 13. If you "transposed" this conditional it would be the probability of B given A—for example, the probability a card is a heart, given that it is a queen, which is instead 1 in 4. Conflating the two as equal is the fallacy of the transposed conditional. In the forensic context, the fallacy would be to conflate one conditional probability—"the chance that he would match the crime scene evidence given that he is innocent," with the transposed conditional—"the chance that he is innocent given that he matches the crime scene evidence." Imagine, for example, that there was A positive blood found at a crime scene, and that the arrested suspect has A positive blood (along with 30% of the rest of the population). We might say that the chance a random person would "match" the blood evidence, assuming he is innocent, is 30%. But it's definitely not true that the chance a person is innocent, given the mere fact that he "matches" the blood evidence, is 30%. In a population of a million people, a full 300,000 people would "match" the blood evidence simply by coincidence. The chance that any one of these 300,000 random people is innocent, given that they "match," is certainly more than 30%; indeed, it is overwhelmingly high.

But when the probabilities become very small, people have a hard time with this concept. In the DNA context, the error would be as follows. Imagine a random match probability (say, 1 in a million), which is the probability that, given a person is not the source of the DNA, they will "match" by sheer chance. This means we would expect that every 1 in a million people would have this profile, and that, in a country of 300 million people, around 300 would "match" by chance. The fallacy would be to inaccurately present it as the probability that, given a person "matches" the DNA at a crime scene, there is only a 1 in a million chance they are not the source.[101] In the shoe print context, the error would be similar. If one starts with the statement, "given a randomly selected person, the probability they will have these particular chevron and wave patterns on their shoe is small," the fallacy would be to claim that "given this person had these particular chevron and wave patterns on their shoe, the chance that they are not the source is small."

Reliability Concerns

Concerns about foundational reliability. As reflected in the governmental and NGO reports discussed in the previous subsection titled "Prior Government and NGO Review of Forensic Feature Comparison Evidence," lingering concerns

101. *See, e.g.*, McDaniel v. Brown, 558 U.S. 120 (2010) (noting that the prosecutor and the government's DNA expert both engaged in the fallacy of the transposed conditional in their statements before the jury); Andrea Roth, *Safety in Numbers?: Deciding When DNA Alone Is Enough to Convict*, 85 N.Y.U. L. Rev. 1130 (2010) (explaining the fallacy in laypersons' terms).

exist about the foundational reliability of some non-DNA forensic feature comparison methods. These are important concerns, as "it has become apparent, over the past decade, that faulty forensic feature comparison has led to numerous miscarriages of justice."[102] In brief, critics focus on four main issues.

First, non-DNA methods lack variability data that show how rare the compared features are in the relevant population. Thus, it is difficult to know merely from the presence of shared features between two items how strong the inference should be that the two items share a common source.[103] While some examiners might rely on their own casework experience to opine about the likely frequency of features in the relevant population, the 2016 PCAST Report cautioned against reliance on examiner experience on this front without empirical data.[104]

Second, these methods are relatively subjective compared to DNA.[105] Because they are based largely on individual examiner judgment and experience, the feature comparisons themselves are sometimes critiqued as more prone to contextual bias and less able to be repeated and reproduced both by the same examiner and by other examiners, which are hallmarks of good science (reproducibility and repeatability).

Third, critics argue there are too few well-designed validation studies indicating the extent of a method's accuracy. Put differently, these methods have not been sufficiently shown through well-designed empirical testing to consistently get the right answer. Regardless of how subjective or black box–like a method is, the method can still be tested against a known truth to determine how often the

102. 2016 PCAST Report, *supra* note 20, at 44 (citing exoneration compilations); *see also* Kori Khan & Alicia Carriquiry, *Hierarchical Bayesian Non-response Models for Error Rates in Forensic Black-Box Studies*, 381 Phil. Transactions Royal Soc'y A: Mathematical, Physical, & Eng'g Scis. 1, 2 (2023), https://doi.org/10.1098/rsta.2022.0157 (explaining that a primary cause of these wrongful convictions is the "ad hoc, subjective methods to evaluate evidence, and the exaggerated claims by expert witnesses at trial").

103. *See, e.g.*, 2009 NRC Report, *supra* note 37, at 149 (critiquing lack of variability studies in non-DNA methods); Robin Mejia et al., *What Does a Match Mean? A Framework for Understanding Forensic Comparisons*, 16 Significance 25–28 (2019), https://doi.org/10.1111/j.1740-9713.2019.01251.x (noting the lack of population data to explain the rarity of characteristics and the resulting inability to accurately determine the probative value of a match).

104. 2016 PCAST Report, *supra* note 20, at 55 ("The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of 'judgment.' It is an empirical matter for which only empirical evidence is relevant.").

105. *See, e.g.*, *id.* at 47 ("Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure of similarity) are obviously more susceptible to human error, bias, and performance variability across examiners."). For more on the need for objective measures, see Karen Kafadar, *The Need for Objective Measures in Forensic Science*, 16 Significance 16 (2019), https://doi.org/10.1111/j.1740-9713.2019.01249.x. Of course, DNA analysis itself is also subjective in many respects. *See generally* Erin Murphy, *The Art in the Science of DNA: A Layperson's Guide to the Subjectivity Inherent in Forensic DNA Typing*, 58 Emory L.J. 489 (2008).

method gets the right answer.[106] Such testing can be used to calculate an error rate, giving factfinders a better sense of how strongly the evidence supports the inference the examiner has drawn. The 2023 amendments to Rule 702 reflect the growing consensus that courts must be willing to "receive an estimate" of the error rates on methodologies employed. The advisory committee's notes' emphasis on error rates is consistent with the independent consensus positions of multiple national committees[107] as reflected in the 2009 NRC Report,[108] the 2016 PCAST Report,[109] and other recent reports from scientific review boards and experts.[110]

The 2016 PCAST Report in particular concluded that non-DNA feature comparison specialties vary widely in terms of well-designed validation studies. The report, after reviewing existing validation studies for a variety of disciplines, concluded that friction ridge analysis (fingerprint comparison) and analysis by expert systems of certain DNA mixtures have proof of foundational reliability with a known error rate,[111] while the validity of other methods (such as microscopic hair comparison, toolmark analysis, and bitemark analysis) had not been

106. *See* 2016 PCAST Report, *supra* note 20, at 5–6 (noting that "black box" methods can and must be tested).

107. *See generally* Paul C. Giannelli, *Forensic Science:* Daubert*'s Failure*, 68 Case W. Rsrv. L. Rev. 869 (2018) (discussing the conclusions of 2009 NRC Report, PCAST, and the NCFS).

108. *See* discussion at section titled "Prior Government and NGO Review of Forensic Feature Comparison Evidence" above.

109. *See id.*

110. *See, e.g.*, Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 Int'l Stat. Rev. 111, 120 (2015), https://doi.org/10.1111/insr.12069 (Expressing support for the 2009 NRC Report and opining that "[t]he scientific method is very general, and the principles of science apply to all branches, irrespective of the specific application (e.g. physics, chemistry or biology). If the scientific method has not been applied, then scientists from no branch can defend the results."); William Thompson et al., Am. Ass'n for the Advancement of Sci. (AAAS), Forensic Science Assessments: A Quality and Gap Analysis: Latent Fingerprint Examination 7–8, 43–44 (2017), https://www.aaas.org/report/latent-fingerprint-examination [hereinafter AAAS Fingerprint Report] (noting the need to quantify uncertainty in the method); Melissa Taylor et al., NIST, Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach (2020), https://doi.org/10.6028/NIST.IR.8282 [hereinafter 2020 NIST Report] (same); Kelly Sauerwein et al., Bitemark Analysis: A NIST Scientific Foundation Review (2023), https://doi.org/10.6028/NIST.IR.8352 [hereinafter NIST Bitemark Report] (same).

111. 2016 PCAST Report, *supra* note 20, at 9 ("PCAST finds that latent-fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis."). *See* also PCAST Addendum, *supra* note 30, at 4:

> [T]he friction-ridge discipline . . . has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful "white-box" studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method.

supported by even one appropriately designed validation study.[112] More recent studies since the 2016 PCAST Report have echoed these concerns.[113]

Fourth, some examiners have used methods for a purpose or context beyond its scientific foundations. Even if a method might be foundationally valid for one modest purpose (such as determining whether two projectiles were fired from the same brand and model of firearm), it might not be foundationally valid for a more ambitious purpose (such as determining the likelihood that two projectiles were fired from the same firearm). Similarly, a handwriting comparison method might be foundationally valid for determining whether two exemplars were written by a different person (exclusion), but not for determining the likelihood that they were written by the same person. The more ambitious an examiner's claim on the witness stand, the more the concern that the method has been stretched beyond its scientific foundations. This could also be an issue with the method's reliability as applied, but it could be seen as an issue of foundational reliability if the method itself is simply not fit for the purpose it is used for.

These continuing issues with non-DNA feature comparison methods are likely in part a reflection of their origins outside the derivative sciences. For the most part, non-DNA feature comparison methods did not develop as "scientific" methods; rather, they grew from forensic practices that did not have governing

112. *See* PCAST Addendum, *supra* note 30, at 5 ("In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitemark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method."); *id.* at 6 (With respect to non-DNA hair comparison, "the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a forensic feature-comparison method, hair comparison lacks a scientific foundation. . . . PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method."); NIST Bitemark Report, *supra* note 110, at 24 (concluding "forensic bitemark analysis lacks a sufficient scientific foundation" and stating that "the data available does not support the accurate use of bitemark analysis to exclude or not exclude individuals as the source of the bitemark").

113. A 2018 study concluded that the method of fingerprint experts is "far from error free." *See* Garrett & Mitchell, *supra* note 19, at 908 (summarizing their findings of an error rate ranging from 1–2% to 10–20% per test with respect to false positives, with an average of 7% false positives and 7% false negatives over a nearly 20-year period from 1995–2016). Several black box studies have been conducted in multiple forensic disciplines since the 2016 PCAST Report, suggesting a very low rate of error. However, those examining the studies claim that the accuracy is lower than reported, owing to a variety of causes that reflect "missingness problems": self-selected participants, high rates of attrition, nonresponse, how inconclusive responses are used, low reproducibility (inter-examiner consistence) and repeatability (intra-examiner consistency). Khan & Carriquiry, *supra* note 102, at 3–6. The "tremendously high rates of missingness" precludes a true estimate of error rates. *Id.* at 17–18. For a critical analysis of the error rates of black box studies conducted on firearms analysis, see Heike Hofmann, Alicia Carriquiry & Susan Vanderplas, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 L., Probability & Risk 317, 342–43 (2020), https://doi.org/10.1093/lpr/mgab002 (addressing the effect of inconclusive findings on error rate accuracy and concluding that there is "significant work to be done" before the authors could confidently provide an error rate associated with firearms and toolmark analysis).

standards, empirical testing, or controlled methods.[114] There were few, if any, scientists involved in the development of these specialties.[115] Many in the forensic community were resistant to change, and historically courts were lenient in evaluating admissibility of such evidence.[116]

While courts frequently admit feature comparison evidence without limitation in many specialties,[117] several federal courts have noted these concerns, particularly in the areas of handwriting comparison and firearm comparison.[118] Even with respect to fingerprint analysis, Judge Frank Easterbrook, writing for the U.S. Court of Appeals for the Seventh Circuit, noted that the method's "false positive rate . . . is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis."[119] Following the suggestions of the advisory committee's notes to the 2023 amendment to Rule 702, courts should ensure that examiner testimony is limited to what is supportable by the empirical data. Courts might choose, for example, to limit the direct testimony to the known error rate, restrict experts' certainty of conclusions, limit the language experts use, or in some cases exclude the evidence entirely.[120]

Concerns about reliability as applied. According to PCAST and other commentators, one key to showing a feature or pattern method's "reliability as applied" is proficiency testing, or "ongoing empirical tests to 'evaluate the capability and performance of analysts.'"[121] Such testing is particularly critical where a method rests in large part on the expert's human judgment, which is especially vulnerable to error and inter-examiner variability.[122] Given how central an

114. *See generally* Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 U.C.L.A. L. Rev. 725 (2011) (noting the nonscientific origin of pattern disciplines and the lack of a culture of scientific inquiry in these fields).

115. Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 Science 892, 893 (2005), https://doi.org/10.1126/science.1111565.

116. 2009 NRC Report, *supra* note 37, at 108–09 (citations omitted).

117. Giannelli, *supra* note 107 (explaining that, with few exceptions, courts have continued to admit forensic science in the same way as they did pre-*Daubert*).

118. *See* sections titled "Handwriting Evidence" and "Firearms and Toolmark Analysis" below.

119. United States v. Bonds, 922 F.3d 343, 345 (7th Cir. 2019).

120. "Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error." Fed. R. Evid. 702 advisory committee's note to 2023 amendment. The notes additionally urge that when making the decision whether to admit forensic expert testimony, "the judge should (where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results." *Id.*

121. 2016 PCAST Report, *supra* note 20, at 57.

122. *Id.* at 58. *See also* Garrett & Mitchell, *supra* note 19, at 902 (explaining why "proficiency testing is the only objective means of assessing the accuracy and reliability of experts who rely on subjective judgments to formulate their opinions (so-called 'black-box experts')").

expert's individual skill is to the success (or not) of relatively subjective methods, the demonstration of such individual skill is important, and—according to these commenters—the only way to meaningfully demonstrate such skill is by measuring each expert's successful performance on realistic proficiency tests.[123] One recurring challenge, however, is that existing proficiency tests tend to be relatively easy, arguably not a meaningful check on examiner competence.[124]

Of course, proficiency testing alone only shows that the expert is qualified, not that the expert reliably applied the method in the case at hand. Also relevant to showing reliability as applied is documentation that the method was correctly applied in the case at hand, by the particular laboratory and examiner(s). This documentation can come in the form of internal validation, protocols, bench notes, reports, and testimony as to what the expert did and did not do in the particular case and the conditions of the sample and testing.

Moreover, to stay within the bounds of validity as applied, the examiners cannot make claims that go beyond the empirical evidence, and any limitations of the method should be acknowledged along with the demonstrated error rate of the method used.[125] The advisory committee's note to the 2023 amendment to Rule 702(d) makes clear that the expert's application of the methodology should be deemed unreliable if they overstate the strength of the inference that can be drawn from the methodology's application in a given case.[126] Expert opinion testimony regarding the weight of feature comparison evidence (i.e., evidence that a set of features corresponds between two examined items) must be limited to those inferences that can reasonably be drawn from a reliable application of the principles and methods. Thus, a statement of certainty or source attribution might not be a reliable application of a particular method, even if a statement about a class characteristic is.

To be sure, most courts are hesitant to exclude testimony where there are concerns about how well the examiner has applied the methodology to the case at hand, often finding such issues are properly matters for cross-examination.[127]

123. Given the wide variability of accuracy among examiners, each expert should be performance-tested. *See* Bradford T. Ulery et al., *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 Proc. Nat'l Acad. Scis. 7733 (2011), https://doi.org/10.1073/pnas.1018707108 (noting the disagreement among examiners about whether fingerprints were suitable for reaching a conclusion).

124. The president of the Collaborative Testing Services (CTS) has candidly acknowledged that "[e]asy tests are favored by the community" of customers for such tests. 2016 PCAST Report, *supra* note 20, at 57 n.133 (quoting president of CTS).

125. *Id*. at 50–51. Note that calculation of the error rate might itself be controversial in a particular case, given issues related to, for example, whether to treat inconclusive determinations as false positives. *See* discussion at section titled "Firearms and Toolmark Analysis" below (discussing firearms validation studies).

126. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

127. *See, e.g.*, Itiel E. Dror, Bridget M. McCormack & Jules Epstein, *Cognitive Bias and Its Impact on Expert Witnesses and the Court,* 54 Judges J. 8, 11 (2015) (the prevailing view is that "errors

Nonetheless, there are decisions where courts have not permitted testimony where the practitioner has not reliably applied the methodology to the facts of the case.[128] In any event, the 2023 amendments to Rule 702 make clear that testimony is not admissible where the expert has not reliably applied the methodology to the case at hand. This shortcoming is not simply a matter for cross-examination but is critical to the decision about admissibility. If the judge is not satisfied by a preponderance of the evidence that the expert has reliably applied the methodology to the case at hand, the testimony should not be admitted.

Discussion of Selected Forensic Feature Comparison Evidence

Differences Between DNA and Other Methods

This reference guide addresses only non-DNA forensic feature comparison techniques, in part because DNA is ubiquitous and complex enough that it is deserving of separate treatment, and in part because of inherent differences between DNA and other existing techniques. Judges would benefit from a basic understanding of these differences.

First, forensic DNA typing was developed in the medical research context. As discussed above, most forensic disciplines were developed by law enforcement, housed in law enforcement facilities, and staffed by law enforcement personnel. In the 2009 NRC Report's view, this lack of independence from law enforcement has contributed to contextual bias, insistence upon a "zero error" rate and other unscientific claims, and a dearth of empirical studies.[129] Whatever one thinks of the NRC's conclusions in this regard, the fact is that much of the literature judges will read in forensic science litigation will tout this difference as a reason to view single-source DNA as the "gold standard" of forensic feature comparison. Second, the criteria for determining whether two DNA profiles are consistent with each other are relatively objective, precise, and repeatable compared to other feature comparison methods. While there is certainly some

in application should result in the exclusion of evidence only if they render the expert's conclusions unreliable; otherwise, the jury should be allowed to consider whether the expert properly applied the methodology in determining the weight or credibility of the expert testimony" (quoting State v. Bernstein, 349 P.3d 200, 201 (Ariz. 2015))).

128. *See, e.g.*, State v. McPhaul, 808 S.E.2d 294, 305 (N.C. Ct. App. 2017) (holding that the trial court abused its discretion in admitting fingerprint comparison evidence where the witness failed to explain how she reached her conclusion in the case at hand).

129. *See* 2009 NRC Report, *supra* note 37, at 71, 99, 104. *See generally* Erin E. Murphy, *What "Strengthening Forensic Science" Today Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report*, 9 L., Probability & Risk 7 (2010), https://doi.org/10.1093/lpr/mgp030 (explaining the NRC's comparison of DNA typing to other more subjective techniques).

judgment involved in determining what is and is not a "real" genetic marker, it is more straightforward to compare two graphs and note that both have a particular marker at a particular location than it is to determine that a smudged characteristic in a latent fingerprint "matches" a characteristic in a reference print. Moreover, as the *Reference Guide on Human DNA Identification Evidence*, in this manual, discusses, validation studies estimate the variability of genetic markers in the population. Third, the data underlying DNA testing allow for a quantified measure of uncertainty rather than merely a statement that two features or patterns "match" or are "consistent," or reliance on an examiner's judgment to decide how many similarities merit a source identification.

Ultimately, DNA will always be different from other feature comparison disciplines. DNA has high probative value in part because it is based on a biological model that explains inheritance; a statistical model that faithfully corresponds to the biology; and the population data that enables estimation of the statistical model parameters. This unique paradigm for DNA cannot be applied in the same way to other types of evidence (such as friction ridge examinations), much less evidence that is manufactured (like a toolmark). That said, other aspects of DNA typing, such as collection of population data about the variability and frequency of features and development of quantified match statistics and error rates, could be replicated in non-DNA disciplines with more research and testing.

Fingerprint Evidence

Introduction

Fingerprinting has been a forensic technique since the mid-1800s,[130] and English and American courts have accepted fingerprint identification testimony for just over a century.[131] Over the years, at least before DNA, fingerprint analysis

130. Francis Galton authored the first textbook on the subject. Francis Galton, Fingerprints (1892) (describing the history of forensic fingerprinting, the ridges on hands and feet, and suggesting techniques for measuring and comparing ridges, including alleged racial differences). The origin of fingerprinting, like that of some other biometric techniques, is fraught in its connections to eugenics and "scientific racism," and is far different from the discipline's modern uses. *See generally* Simon Cole, Suspect Identities: A History of Fingerprint and Criminal Identification (2001). For more on the history of fingerprint identification, *see* Jennifer L. Mnookin, *Fingerprint Evidence in an Age of DNA Profiling*, 67 Brook. L. Rev. 13 (2001).

131. United States v. Llera Plaza, 188 F. Supp. 2d 549, 572 (E.D. Pa. 2002) (describing a 1906 English case in which "a New York City detective who had in 1904 been posted to Scotland Yard to learn about fingerprinting . . . used his new training to break open two celebrated cases: in each instance fingerprint identification led the suspect to confess. . . ."). The first published American appellate opinion sustaining admission of fingerprint testimony was *People v. Jennings*, 96 N.E. 1077 (Ill. 1911).

became known as the "gold standard" of forensic feature comparison expertise.[132] In fact, proponents of new, emerging techniques in forensics would sometimes attempt to invoke onto the new techniques the prestige of fingerprint analysis.[133] Likewise, some early proponents of DNA typing alluded to it as "DNA fingerprinting."[134] Fingerprint comparison is still widely used and is one of the most relied-upon forms of forensic feature comparison across the world.

This is not to say that fingerprint comparisons are free from error. In fact, the FBI's incorrect identification of attorney Brandon Mayfield as the perpetrator of the 2004 Madrid train bombing is perhaps the most well-known error and one that spurred additional research.[135] Like other forms of forensic feature comparison, fingerprint comparison was judicially adopted without any proof of its foundational reliability. As noted below, at least one court has limited expert testimony about fingerprints and government reports have raised new questions about the limits of its accuracy.

The Method and Its Claims

Fingerprint analysis involves comparing the features of an evidence print (typically a "latent" invisible print lifted from a surface) and a reference print to determine if they might have a common source. Fingerprint comparison has historically been based on three assumptions: (1) the uniqueness of each person's friction ridges on their fingers, (2) the permanence of those ridges throughout a person's life, and (3) the transferability of an impression of that uniqueness to another surface. Contemporary research addresses each of these assumptions. First, scientific research has "convincingly established" that ridge patterns of humans' fingers vary greatly among individuals and, as such, fingerprint comparison is a theoretically viable way to distinguish individuals.[136] Second, fingerprints appear to be persistent over one's lifetime, although there can be minor changes in friction ridge detail due to aging or occupation.[137] Third, there is

132. *See* Sandy L. Zabell, *Fingerprint Evidence*, 13 J. L. & Pol'y 143 (2005) (noting fingerprint evidence was long considered the "gold standard" of human identification).

133. *See, e.g.*, Kenneth Thomas, *Voiceprint—Myth or Miracle*, *in* Scientific and Expert Evidence 1015 (2d ed. 2011) (noting that advocates of sound spectrography referred to it as "voiceprint" analysis).

134. Colin Norman, *Maine Case Deals Blow to DNA Fingerprinting,* 246 Science 1556 (1989), https://doi.org/10.1126/science.2688090.

135. *See* Robert B. Stacey, *Report on the Erroneous Fingerprint Individualization in the Madrid Train Bombing Case*, Fed. Bureau of Investigation (Jan. 2005), https://perma.cc/5WY6-AWHE.

136. *See* AAAS Fingerprint Report, *supra* note 110, at 17. This variability extends to related individuals, and even identical twins develop distinguishable friction ridge skill detail. *Id.* (citations omitted). Distinguishing identical twins occurs at a slightly lower accuracy than non-twins, albeit with "relatively high accuracy." *Id.*

137. *Id.* at 27–28 (citing numerous sources).

ample evidence that fingerprints left on a surface can be accurately transferred to another surface for purposes of comparison using various methods, although as stated, the quality (and thus analytical value) of these prints may vary dramatically from "rolled" prints ("record" prints that are typically rolled onto a fingerprint card or digitized and scanned into an electronic file).[138] For example, latent prints from a crime scene vary in pressure, and the elasticity of skin naturally distorts the impression.[139] Consequently, fingerprint impressions from the same person typically differ in some respects each time the impression is left on an object.[140] The latent print might sometimes be so fragmentary or smudged that analysis is impossible.

Fingerprint examiners generally follow a procedure known as analysis, comparison, evaluation, and verification (ACE-V). In the analysis stage, the examiner studies the evidence print (typically a latent print from a crime scene) to determine whether the quantity and quality of details in the print are sufficient to permit further evaluation, and determines what features are present.[141] In the comparison stage, the examiner visually compares the evidence and known prints to determine "the details that correspond" between them, such as the "overall shape" of the prints, "lengths of the ridges, minutia location and type, thickness of the ridges and furrows, shapes of the ridges, pore position, crease patterns and shapes, scar shapes, and temporary feature shapes (e.g., a wart)."[142] In the evaluation stage, the examiner "evaluates the sufficiency of the detail present to establish an identification (source determination)," determining whether, "based on his or her experience," a "sufficient quantity and quality of friction ridge detail is in agreement between the latent print and the known

138. *See* David A. Stoney, *Scientific Status*, *in* Modern Scientific Evidence: The Law and Science of Expert Testimony § 32.29 (David L. Faigman et al., 2021–2022 ed.) (discussing the various methods of extraction); United State v. Cruz-Mercedes, 379 F. Supp. 3d 24, 44–45 (D. Mass.), *aff'd on other grounds*, 945 F.3d 569 (1st Cir. 2019) (explaining how there was only one available latent print from which to make a comparison; the remaining prints were insufficiently clear to be useful).

139. *See, e.g.*, Stoney, *supra* note 138, § 32.32 ("Smudging, dirty surfaces, dirty fingers, and contingencies of fingermark deposition all contribute to the incomplete transfer of the finger's detail and the introduction of specious or artifactual detail."); United States v. Mitchell, 365 F.3d 215, 220–21 (3d Cir. 2004) ("Criminals generally do not leave behind full fingerprints on clean, flat surfaces. Rather, they leave fragments that are often distorted or marred by artifacts. . . . Testimony at the *Daubert* hearing suggested that the typical latent print is a fraction—perhaps 1/5th—of the size of a full fingerprint."); *id.* at 221 n.1 ("In the jargon, artifacts are generally small amounts of dirt or grease that masquerade as parts of the ridge impressions seen in a fingerprint, while distortions are produced by smudging or too much pressure in making the print, which tends to flatten the ridges on the finger and obscure their detail.").

140. 2009 NRC Report, *supra* note 37, at 144 ("The impression left by a given finger will differ every time, because of inevitable variations in pressure, which change the degree of contact between each part of the ridge structure and the impression medium.").

141. *Id.* at 137–38.

142. *Id.* at 138–39. *See also* Stoney, *supra* note 138, § 32.32.

print" to tell whether the prints do or do not come from the same source.[143] In the *verification* stage, a second examiner repeats the analysis to see if they arrive at the same conclusion. Verification is blind if the second examiner does not know the first examiner's conclusion.

Each of these stages in ACE-V affords examiners significant discretion. For example, examiners themselves (at least those in the United States) determine based on their judgment and experience not only how many points of comparison are sufficient before drawing a conclusion,[144] but whether any apparent dissimilarities (which would otherwise require reporting an "exclusion") can be discounted as an artifact or are a result of distortion.[145] The examiner also has broad discretion in how to evaluate such factors as "inevitable variations" in pressure, but to date these factors have not been "characterized, quantified, or compared."[146]

Examiners often use the Automated Fingerprint Identification System (AFIS) of databases to rapidly search fingerprint databases in no-suspect cases. This system of databases was developed in the late 1970s and has become more important for law enforcement with the growth of databases of known prints.[147] AFIS is highly accurate for comparison of all ten rolled or slapped prints to another set of known prints. AFIS searches are also useful for screening large numbers of prints, which can then be evaluated by human examiners.

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of fingerprint analysis, given that these will frame the litigation judges will face over admission of latent print examiner testimony.

Since the 2004 Brandon Mayfield false positive FBI incident and the 2009 NRC Report, critics have focused primarily on the following concerns with

143. 2009 NRC Report, *supra* note 37, at 138.

144. Stoney, *supra* note 138, § 32.35; United States v. Llera Plaza, 188 F. Supp. 2d 549, 566–71 (E.D. Pa. 2002) (noting that U.S. and Scotland Yard examiners have no minimum or set number of points of comparison); 2009 NRC Report, *supra* note 37, at 141 ("The latent print community in the United States has eschewed numerical scores and corresponding thresholds" and consequently relies "on primarily subjective criteria" in making the ultimate attribution decision).

145. Commonwealth v. Patterson, 840 N.E.2d 12, 17 (Mass. 2005):

> There is a rule of examination, the "one-discrepancy" rule, that provides that a nonidentification finding should be made if a single discrepancy exists. However, the examiner has the discretion to ignore a possible discrepancy if he concludes, based on his experience and the application of various factors, that the discrepancy might have been caused by distortions of the fingerprint at the time it was made or at the time it was collected.

146. 2009 NRC Report, *supra* note 37, at 144.

147. AAAS Fingerprint Report, *supra* note 110, at 29. AFIS includes a few million subjects for state law enforcement agencies to over a hundred million for national agencies.

respect to fingerprint analysis: its relative subjectivity, both in terms of reliance on judgment and lack of standards related to determining minimum number of points of comparison or sufficiency of detail (and thus its tendency to allow contextual bias and inter- and intra-examiner variability); the lack of variability data (about rarity of various features) to support determinations of source identification; the lack of well-designed validation studies and accurate error rate estimates; and problems with specific applications of ACE-V, such as overreliance on AFIS, failure to examine a sufficient number of details, problems with certain latent prints that go beyond the method's capabilities, lack of well-designed proficiency testing, and nonblind verification (where the verifying examiner already knows what conclusion the first examiner came to, and thus might be biased in favor of affirming the first conclusion).

Both the FBI's post-Mayfield internal investigative report and the 2009 NRC Report focused on concerns related to subjectivity, lack of variability data, and reliability as applied. The FBI noted that the Mayfield false positive result was likely the result of examiners falsely assuming that the source of the latent print was among the top potential matches suggested by AFIS, and the fact that Mayfield's print and the latent print were, in fact, remarkably similar (and thus also remarkably similar to the real source, a different suspect in Spain).[148] The 2009 NRC Report focused on the relative subjectivity of the method and lack of data as to how rare or common a particular type of fingerprint characteristic is.[149] As the 2009 NRC Report explained, "the ACE-V method does not specify particular measurements or a standard test protocol, . . . examiners must make subjective assessments throughout."[150] Or, as statistician Sandy Zabell put it a year after the Mayfield incident: "In contrast to the scientifically-based statistical calculations performed by a forensic scientist in analyzing DNA profile frequencies, each fingerprint examiner renders an opinion as to the similarity of friction ridge detail based on his subjective judgment."[151] A later study by psychologist Itiel Dror and others found significant differences in latent print examiner conclusions related to the same set of samples depending on whether the examiner had been exposed to task-irrelevant information, and inconsistency within the same examiner's conclusions over time.[152] The 2009 NRC Report's conclusion was that the ACE-V method, while then nearly universally accepted by courts, was too "broadly stated" to "qualify as a validated method for this type of analysis."[153]

148. *See generally* Off. of the Inspector Gen., U.S. Dep't of Justice, A Review of the FBI's Handling of the Brandon Mayfield Case (Mar. 2006), https://perma.cc/7UY6-575D.

149. 2009 NRC Report, *supra* note 37, at 139–40, 144.

150. *Id.* at 139.

151. Zabell, *supra* note 132, at 158.

152. Dror et al., *supra* note 51.

153. 2009 NRC Report, *supra* note 37, at 142.

The 2016 PCAST Report's evaluation of ACE-V focused largely on (1) whether sufficient well-designed validation studies existed and what the estimated error rate was; and (2) the reliability of ACE-V as applied. The PCAST Report identified two friction ridge studies that met its criteria for a well-designed validation study, meaning one with samples representative of real casework; a large enough sample to generate a meaningful error rate estimate; double-blind (meaning subject and grader do not realize it is a test or what the results are); and not subject to changing testing protocol midstream.[154] First, the FBI published a peer-reviewed study, in 2011, that was conducted in direct response to the 2009 NRC Report.[155] The false positive rate from the FBI study was 0.17%, with an upper bound of a 95% confidence interval of 0.33% (meaning, in simple terms, that the error rate is likely to be within a certain range with 0.33% being the highest in the range).[156] In terms of application, these rates correspond to a false positive occurring in an estimated 1 in every 604 cases, but potentially as often as 1 in every 306 cases.[157] In 2012, the FBI conducted a follow-up study directed toward assessing repeatability and reproducibility. In the follow-up study, 75 of the original 169 examiners participated, and the false positive rate was broadly consistent with the prior study.[158]

The second and final ACE-V study identified by the PCAST Report as meeting its criteria for a well-designed validation study was done in 2014 by the Miami-Dade Police Department Forensic Services Bureau.[159] This study had a false positive rate of 4.2%, with an upper bound of a 95% confidence interval of 5.4%.[160] These rates correspond to a false positive occurring in an estimated 1 in 24 cases, but potentially as often as 1 in 18 cases.[161] Of note, the study's authors identified 35 potential clerical errors in the documentation of the participants.[162] If the 35 known errors were properly accounted for, then the actual false positive rate would have been 0.7%, with an upper bound of a 95% confidence interval of 1.4%.[163] These rates would correspond to a false positive occurring in an estimated 1 in 146

154. 2016 PCAST Report, *supra* note 20, at 94–97 (two studies); 52 (noting criteria for well-designed study).

155. *Id.* at 94.

156. *Id.*

157. *Id.*

158. *Id.* For more on the repeatability and reproducibility of examiner's decisions, see Bradford T. Ulery et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*, 7 PLoS ONE e32800 (2012), https://doi.org/10.1371/journal.pone.0032800.

159. 2016 PCAST Report, *supra* note 20, at 94–95. The study can be found at https://perma.cc/C47K-GCQA.

160. *Id.* at 95.

161. *Id.*

162. *Id.*

163. *Id.*

cases, but potentially as often as 1 in 73 cases.[164] However, the report noted that to exclude errors in a post hoc manner was inappropriate for a validation study.[165]

The 2016 PCAST Report ultimately concluded based on the two well-designed studies that "latent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis."[166] Note that the report's use of the term foundationally valid appears to track the concept of "foundational reliability" in *Daubert* and its progeny. The 2016 PCAST Report did also suggest that juries in cases involving latent print expert testimony be informed that: (1) there have only been two properly conducted studies whose results are worth considering; and (2) the false positive rates from these studies "could be as high as 1 in 306 in one study and 1 in 18 in the other study."[167] It also suggested that examiners acknowledge that because the participants knew they were participating in the study, the actual false positive rate could be higher than observed.[168] Regarding this third point, however, the report also noted, in a separate section, that "[i]t is likely that a properly designed program of systemic, blind verification would decrease the false-positive rate, because examiners in the studies tend to make *different* mistakes."[169] Giving this information to factfinders, according to the report's authors, would allow jurors "to weigh the probative value of the evidence."[170]

Even though it concluded latent fingerprint analysis is foundationally valid (or foundationally "reliable" in *Daubert* parlance), the 2016 PCAST Report concluded there were several issues related to its validity (or reliability, in *Daubert* parlance) *as applied*.[171] These issues are: (1) confirmation bias—examiners altering initially marked features in a latent print based on comparison with an exemplar, (2) contextual bias—other facts of the case influencing the examiner's judgment, and (3) lack of proficiency testing.[172] Given these concerns, the report concluded the following would be necessary to establish validity of ACE-V as applied:

> From a scientific standpoint, validity as applied requires that an expert: (1) has undergone appropriate proficiency testing to ensure that he or she is

164. *Id.*

165. *Id.*

166. *Id.* at 101.

167. *Id.* at 96. The 2016 PCAST Report authors determined the number of studies worth considering based on their framework; other studies may provide helpful information. Moreover, the report authors urged that the error rates from existing validation studies be calculated based on the number of incorrect calls as a percentage of all *conclusive* examinations, and that they treat "inconclusives" as false positives. *Id.* at 93.

168. *Id.* at 109.

169. *Id.* at 96 (emphasis in original).

170. *Id.*

171. *Id.* at 102.

172. *Id.*

> capable of analyzing the full range of latent fingerprints encountered in casework and reports the results of the proficiency testing; (2) discloses whether he or she documented the features in the latent print in writing before comparing it to the known print; (3) provides a written analysis explaining the selection and comparison of the features; (4) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion; and (5) verifies that the latent print in the case at hand is similar in quality to the range of latent prints considered in the foundational studies.[173]

Note that while proficiency testing might also be related to the qualification of a latent print examiner, the requirements above, beyond proficiency testing alone, relate to documentation of the method sufficient to support a finding by a preponderance of the evidence that the examiner reliably applied ACE-V in a particular case.

The findings of PCAST as to ACE-V's error rate and limitations were repeated in a report on latent print examination the following year by the AAAS.[174] One difference, however, was that the AAAS report looked at the entire body of research surrounding latent fingerprint analysis, whereas the PCAST report drew its conclusions from what it determined were the two appropriately designed black box studies.[175] Despite the differences in scope of review, AAAS stated that its "conclusions largely align with those of the PCAST report."[176]

The AAAS report also mirrored the 2009 NRC Report's concern about lack of empirical data underlying source conclusion and concluded that statements of source attribution are as yet scientifically unsupportable. While the AAAS report concluded that research had "convincingly established" that ridge patterns vary greatly among humans,[177] it found no scientific basis for determining the rarity of any such ridge feature[178] or "how many features, of what types, are needed in order for an examiner to draw definitive conclusions about whether a latent print was made by a given individual."[179] As a result, "[e]xaminers may well be able to exclude the preponderance of the human population as possible sources of a latent print, but there is no scientific basis for estimating the number of people who could not be excluded," and there is no science to support the determination that the latent print came from a single source.[180] Thus, any notion that an examiner can make or has

173. *Id.*
174. AAAS Fingerprint Report, *supra* note 110.
175. *Id.*
176. *Id.* at 44.
177. *Id.* at 23.
178. *Id.*
179. *Id.* at 21.
180. *Id.*

made an identification with 100% accuracy is "overstated" and "indefensible,"[181] as is any "claim that they can associate a latent print with a single source."[182]

The AAAS report then offered several suggestions to address any misconceptions a juror may harbor based on previous assertions of either source attribution or infallibility:

> [L]atent print examiners should include specific caveats in reports that acknowledge the limitations of the discipline. They should acknowledge: (1) that the conclusions being reported are opinions rather than facts (as in all pattern-matching disciplines); (2) that it is not possible for a latent print examiner to determine that two friction ridge impressions originated from the same source to the exclusion of all others; and (3) that errors have occurred in studies of the accuracy of latent print examination.[183]

The AAAS report also recommended that examiners refrain from making statements that are not supported by the science behind latent fingerprint examination.[184] Any words or phrases—such as "match," "identification," and "individualization"—that would imply a single source should be avoided.[185] The AAAS report recommended proposed testimony an examiner might offer that avoids any such peril:

> The latent print on Exhibit ## and the record fingerprint bearing the name XXXX have a great deal of corresponding ridge detail with no differences that would indicate they were made by different fingers. There is no way to determine how many other people might have a finger with a corresponding set of ridge features, but this degree of similarity is far greater than I have ever seen in non-matched comparisons.[186]

The Department of Justice, for its part, has changed its source conclusion language and other practices (such as nonblind verification) in response to the reports cited above. The FBI, for example, now forbids claims of "individualization."[187] DOJ's Uniform Language for Testimony and Reports (ULTR) also prohibits testifying examiners from "assert[ing] that a 'source identification' or a 'source exclusion' conclusion is based on the 'uniqueness' of an item of evidence," "us[ing] the terms 'individualize' or 'individualization' when describing a source conclusion," or "assert[ing] that two friction ridge skin impressions originated from the same

181. *Id.* at 71.

182. *Id.* at 60. U.S. Dep't of Justice, Approved Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline (2018), https://perma.cc/E59D-B6C2.

183. AAAS Fingerprint Report, *supra* note 110, at 73.

184. *Id.* at 11.

185. *Id.*

186. *Id.*

187. *See* Fed. Bureau of Investigation, FBI Approved Standards for Scientific Testimony and Report Language for Forensic Document Comparisons 4 (2022), https://perma.cc/VZ5C-5ZBT.

source to the exclusion of all other sources."[188] These prohibitions are in addition to the proscription of asserting latent fingerprint analysis is "infallible or has a zero error rate."[189]

Notwithstanding the recent purging of terms like individualization and zero error rate, nearly all laboratories, including DOJ's, still allow examiners to testify to language suggesting categorical source attribution or exclusion. DOJ's ULTR now directs their examiners to arrive at one of three conclusions: a source identification, a source exclusion, or an inconclusive determination.[190] A source identification is an examiner's opinion that the two prints came from the same person and is based on an examiner's opinion that there is "extremely strong support" indicating the two prints came from the same source and "extremely weak support" indicating the two prints came from different sources.[191] A source exclusion is an examiner's opinion that the two prints did not come from the same source and is based on the examiner's opinion that there is "extremely strong support" indicating the two prints came from different sources and "extremely weak or no support" that the two prints came from the same source.[192] An examiner will make an inconclusive determination when "there is insufficient quantity and/or clarity" between the two impressions for the examiner to arrive at a source identification or source exclusion determination.[193] While these terms do not claim infallibility, they do suggest the ability to reliably conclude that prints have a single common source. Concerns still linger, of course, about whether jurors understand the significance of such limitations.[194]

A final recurring concern relates to use of AFIS databases. The Mayfield case was one high-profile example of a false hit from a database search. As databases grow, they may contain more fingerprints with many common features and few discernible dissimilar features, known as close non-"matches" (CNMs). In a study evaluating 125 fingerprint agencies completing a mandatory proficiency test with two pairs of CNMs, the false positive rate was 15.9% and 28.1% respectively, raising serious concerns.[195] While there are continued efforts to

188. U.S. Dep't of Justice (DOJ), Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline 3 (2020), https://perma.cc/J7GQ-US5Y.

189. *Id.*

190. *Id.* at 2.

191. *Id.* In other words, "[a]n identification is the statement of an examiner's opinion (an inductive inference) that the probability that the two impressions were made by different sources is so small that it is negligible" (internal citation omitted).

192. *Id.*

193. *Id.* at 3.

194. In a study in the firearms context evaluating juror comprehension about differences in expert's conclusion language, the authors found that more modest phrasing about conclusions did not affect jurors' conclusions. *See* Brandon L. Garrett et al., *Mock Jurors' Evaluation of Firearm Expert Testimony*, 44 Law & Hum. Behav. 412 (2020), https://doi.org/10.1037/lhb0000423.

195. *See* Jonathan J. Koehler & Shiquan Liu, *Fingerprint Error Rate on Close Non-matches*, 66 J. Forensic Scis. 129, 131–32 (2020), https://doi.org/10.1111/1556-4029.14580. The authors

improve AFIS,[196] the systems are not currently designed to prove that a particular pair of prints has a common source.[197]

Case Law Development

Since the Illinois Supreme Court's approval of fingerprint analysis in *People v. Jennings* (1911), fingerprint testimony has been routinely admitted.[198] Indeed, some courts stated that fingerprint evidence was the strongest proof of a person's identity.[199] With the exception of one federal district court decision that was later withdrawn by the court itself upon reconsideration, and another decision in which a court remanded to determine whether ACE-V might be unreliable as applied to simultaneous impressions (rather than single latent prints),[200] nearly all courts before 2017 continued to reject challenges to fingerprint testimony.[201]

note that there were limitations to the study; the participants knew they were being tested and the samples were "convenience samples," not random and representative selections from database searches. For more on AFIS searches, see Itiel E. Dror & Jennifer L. Mnookin, *The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science*, 9 Law, Probability & Risk 47 (2010), https://doi.org/10.1093/lpr/mgp031.

196. AAAS Fingerprint Report, *supra* note 110, at 33, quoting the 2016 PCAST Report, *supra* note 20.

197. AAAS Fingerprint Report, *supra* note 110, at 33.

198. People v. Jennings, 96 N.E. 1077 (Ill. 1911). As Professor Mnookin has noted, "fingerprints were accepted" after *Jenkins* "as an evidentiary tool without a great deal of scrutiny or skepticism." Mnookin, *supra* note 130, at 17.

199. People v. Adamson, 165 P.2d 3, 12 (Cal. 1946), *aff'd*, 332 U.S. 46 (1947).

200. *See, e.g.*, Commonwealth v. Patterson, 840 N.E.2d 12, 15, 16–17 (Mass. 2005) ("These latent print impressions are almost always partial and may be distorted due to less than full, static contact with the object and to debris covering or altering the latent impression"; "In the evaluation stage, . . . the examiner relies on his subjective judgment to determine whether the quality and quantity of those similarities are sufficient to make an identification, an exclusion, or neither."); *and* United States v. Llera Plaza, 179 F. Supp. 2d 492, 516, *vacated, mot. granted on recons.*, 188 F. Supp. 2d 549 (E.D. Pa. 2002). The ruling excluded expert testimony that two sets of prints "matched," to the exclusion of all other persons. On a motion for reconsideration, the court reversed itself. A spate of legal articles followed. *See, e.g.*, Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from* Jennings *to* Llera Plaza *and Back Again*, 41 Am. Crim. L. Rev. 1189 (2004); Robert Epstein, *Fingerprints Meet* Daubert: *The Myth of Fingerprint "Science" Is Revealed*, 75 S. Calif. L. Rev. 605 (2002); Kristin Romandetti, *Recognizing and Responding to a Problem with the Admissibility of Fingerprint Evidence Under* Daubert, 45 Jurimetrics 41 (2004).

201. *See* United States v. Baines, 573 F.3d 979, 990 (10th Cir. 2009) ("Fingerprint identification has been used extensively by law enforcement agencies all over the world for almost a century."); United States v. Abreu, 406 F.3d 1304, 1307 (11th Cir. 2005) ("We agree with the decisions of our sister circuits and hold that the fingerprint evidence admitted in this case satisfied *Daubert*."); United States v. Janis, 387 F.3d 682, 690 (8th Cir. 2004) (finding fingerprint evidence to be reliable); United States v. Mitchell, 365 F.3d 215, 234–52 (3d Cir. 2004); United States v. Crisp, 324 F.3d 261, 268–71 (4th Cir. 2003); United States v. Collins, 340 F.3d 672, 682 (8th Cir.

Since the AAAS Fingerprint Report in 2017, at least one court has limited latent print testimony in new ways on "reliability as applied" grounds. For example, the North Carolina Supreme Court in 2018 held that the trial court abused its discretion in allowing a latent print examiner to testify to a "match" between the defendant's known prints and a latent print on a truck, where the testimony merely established that the examiner used ACE-V, and determined the "match" based on her "training and experience" and "looking at the individual minutia" in each print.[202] The court found a similar error in another case two years later.[203]

Still, since the 2016 PCAST Report, most courts have deemed concerns with latent print analysis as going to weight, not admissibility,[204] citing the long-standing nature of the discipline and its traditional acceptance by courts under *Daubert*.[205] Other courts upholding the admissibility of fingerprint analysis have noted that while reports have exposed flaws in ACE-V, these reports have also opened up new opportunities for defense. One court noted that "[c]ross-examination on issues like error rates is possible now in a way in which it was not 30 years ago."[206] Still other courts have reasoned that even with flaws, fingerprint analysis is likely better than older evidence like eyewitness identifications and "grainy" photographs, suggesting that the middle ground is to admit such evidence and "subject [it] to cross-examination about a method's reliability and whether the witness took appropriate steps to reduce errors."[207]

2003) ("Fingerprint evidence and analysis is generally accepted."); United States v. Hernandez, 299 F.3d 984, 991 (8th Cir. 2002); United States v. Sullivan, 246 F. Supp. 2d 700, 704 (E.D. Ky. 2003); United States v. Martinez-Cintron, 136 F. Supp. 2d 17, 20 (D.P.R. 2001).

202. State v. McPhaul, 808 S.E.2d 294, 304–05 (N.C. Ct. App. 2017), *discretionary review improvidently allowed*, 818 S.E.2d 102 (N.C. 2018) (abuse of discretion to admit where the expert failed to demonstrate that she applied the principles and methods reliably to the facts of the case).

203. *Cf.* State v. Koiyan, 841 S.E.2d 351 (N.C. Ct. App. 2020) (same error as in *McPhaul*, though declining to reverse under plain error review).

204. *See* United States v. Reyes-Ballista, No. CV 18-634-2 (ADC), 2020 WL 6822372, at *4 (D.P.R. Nov. 20, 2020) ("considering that defendant will have ample opportunity to conduct vigorous cross-examination of the government's expert witnesses and present contrary evidence, defendant is not without means of attacking the evidence he now claims to be based on methods that run afoul the profession's parameters and accepted methods").

205. United States v. Pitts, No. 16-CR-550 (DLI), 2018 WL 1116550, at *4–6 (E.D.N.Y. Feb. 26, 2018), citing United States v. Avitia-Guillen, 680 F.3d 1253, 1260 (10th Cir. 2012) ("Fingerprint comparison is a well-established method of identifying persons, and one we have upheld against a *Daubert* challenge."). *See also Reyes-Ballista*, 2020 WL 6822372, at *3 (noting that "several . . . Circuits have explicitly determined that, 'in the context of fingerprint evidence, a *Daubert* hearing is not always required'") (internal citation omitted); United States v. Stevens, 219 F. App'x 108, 109 (2d Cir. 2007) (summary order declining to hold a *Daubert* hearing); United States v. Bonds, No. 15 CR 573-2, 2017 WL 4511061 (N.D. Ill. Oct. 10, 2017), *aff'd*, 922 F.3d 343 (7th Cir. 2019).

206. United States v. Fell, No. 5:01-CRCR-12-01, 2016 WL 11550830, at *1 (D. Vt. Dec. 29, 2016).

207. *Bonds*, 922 F.3d at 346 (Easterbrook, J.).

Given the 2023 amendment to Rule 702, courts may now need to engage in more robust gatekeeping to evaluate both the foundational reliability and the reliability as applied of latent print expert testimony. As the advisory committee's notes clarify, "many courts have held that the critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a)."[208] The notes explain that "[j]udicial gatekeeping is essential." As explained earlier, several gatekeeping options are available to the courts, along with the option of providing jury instructions to aid jurors in evaluating the evidence. With the enactment of the 2023 amendment to Rule 702, the judge must be satisfied that the proponent of the evidence has met each of the rule's requirements by a preponderance. If the court is not satisfied, it should exclude those opinions. Additionally, the judge may appropriately limit any opinion to exclude overstatement of a conclusion.

Handwriting Evidence

Introduction

Individuals who compare handwriting, ink formulations, and other tasks related to evaluation of questioned documents are known as questioned document examiners or forensic document examiners (FDEs). FDEs are called on to perform a variety of tasks such as determining potential authorship of a writing; determining the sequence of strokes on a page; and determining whether a particular ink formulation existed on the purported date of a writing.[209] Courts were originally skeptical of handwriting comparison and dismissed its value, but by 1900, many courts had begun to admit such expert evidence.[210] In the 1936 trial about the Lindbergh baby kidnapping, expert testimony about the handwriting of the random notes figured prominently.[211] Following the Lindbergh

208. Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

209. Questioned document examinations cover a wide range of analyses: handwriting, hand printing, typewriting, mechanical impressions, altered documents, obliterated writing, indented writing, and charred documents. *See* Paul C. Giannelli et al., Scientific Evidence § 14 (6th ed. 2020).

210. *See, e.g.*, D. Michael Risinger, Mark P. Denbeaux, & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. Pa. L. Rev. 731, 762 (1989). For a deep and nuanced analysis of the history of handwriting identification, see Jennifer L. Mnookin, *Scripting Expertise: The History of Handwriting Identification Evidence and the Judicial Construction of Reliability*, 87 Va. L. Rev. 1723 (2001).

211. Risinger et al., *Exorcism of Ignorance*, *supra* note 210, at 771.

case, handwriting evidence became widely used and judicially accepted, but with little critical scrutiny by the courts into its accuracy.[212]

Reliance on handwriting comparison has diminished in recent years and forensic document examination units are closing as demand shifts to other forensic disciplines, such as DNA analysis.[213] Additionally, there is a continual decrease in handwritten communication, as individuals use more electronic communication and digital signatures. These developments suggest that this field may soon join other receding identification specialties. Nevertheless, it continues to be used in court, and in 2020 a NIST working group issued a report, discussed below, for improving the practice.

The Method and Its Claims

Some FDE tasks, such as ink identification, are based on sophisticated techniques such as chromatography and mass spectrometry that are widely considered to be reliable.[214] Other tasks, such as determining authorship of a document, present more substantial potential legal issues. This section will focus primarily on authorship.

Like other feature comparison experts, an FDE determines authorship by analyzing a questioned document and known sample (either previous writing or a requested handwriting exemplar), observing various details in each, comparing them in an attempt to discern consistent or inconsistent patterns, and then evaluating those similarities and differences to arrive at a conclusion about whether they might have a common author. There must be a sufficient quantity and quality of material to permit the examiner to compare the samples.[215] The 2020 NIST Report suggests that it is best if the exemplars are written during the same period as the questioned document, to remove potential variables. In performing their comparison, examiners consider letter formation, size, and inter-word and intra-word spacing.[216]

The so-called conventional or classical approach is a two-stage process. Examiners consider (1) class and (2) so-called "individual" characteristics. Of characteristics labeled "class" characteristics, two types are weighed: "system" and "group."[217] People exhibiting system characteristics would include, for

212. *Id.* (noting that the Lindbergh case seemed to have "stamped out virtually all manifestations of judicial skepticism").

213. 2020 NIST Report, *supra* note 110, at xi.

214. Giannelli et al., *supra* note 209, § 14.04 [4].

215. 2020 NIST Report, *supra* note 110, at 45 (noting the assumption of individuality rests on a large enough quality sample).

216. *Id.* at 14.

217. *See* James A. Kelly, *Questioned Document Examination*, *in* Scientific and Expert Evidence 695, 698 (2d ed. 2011).

example, those who learned the Palmer method of cursive writing, taught in many schools. People taught the same system might reasonably be expected to manifest some of the characteristics of that writing style. However, because schools now devote much less time to teaching handwriting as a skill, the more contemporary view is that the "class" characteristics are not as readily identifiable as they once were.[218] Other "group" characteristics might include those with a medical condition affecting their handwriting.[219]

The conventional belief in individual characteristics unique to a particular person also persists among many examiners. That belief rests on two premises: (1) that no two writers share the same combination of writing characteristics; and (2) adults have consistent writing habits.[220] These beliefs remain unproven,[221] but still assumed by many practitioners and still cited by some courts.[222] The traditional process for handwriting identification involves measuring selected features in handwriting, determining whether and how they differ across specimens, and interpreting the significance of similarities and differences.[223] Although this process can include measuring features, more often it involves an examination of relative measurements (an estimation of features in proportion to each other), including size, spacing, and slant of features. Although there are differences among the various examiners, most follow these general procedures:

1. analyzing the features of the questioned writing and known standards both macroscopically and microscopically;
2. noting conspicuous features such as size, slant, and letter construction, as well as more subtle characteristics such as pen direction, the nature of connections between letters, and spacing between letters, words, and lines;
3. comparing the observed features to determine similarities and dissimilarities;
4. taking into account the degree of similarity or otherwise and the nature of the writing (quality, amount, and complexity), evaluating the evidence,

218. 2020 NIST Report, *supra* note 110, at 7–8.

219. *See, e.g.*, Larry F. Stewart, *The Process of Forensic Handwriting Examinations*, 4 Forensic Res. & Criminology Int'l J. 139 (2017), https://doi.org/10.15406/frcij.2017.04.00126.

220. 2020 NIST Report, *supra* note 110, at 8 (citing Howard Sieden & Frank Norwitch, *Questioned Documents*, *in* Forensic Science: An Introduction to Scientific and Investigative Techniques (S.H. James et al. eds., 4th ed. 2014)).

221. For more, see Andrew Sulner, *Critical Issues Affecting the Reliability and Admissibility of Handwriting Opinion Evidence—How They Have Been Addressed (or Not) Since the 2009 NAS Report, and How They Should Be Addressed Going Forward: A Document Examiner Tells All,* 48 Seton Hall L. Rev. 631, 639 (2018) ("absolutist statements concerning uniqueness and intra-writer variability are as yet unproven, and likely unprovable").

222. *See, e.g.*, United States v. Mallory, 902 F.3d 584 (6th Cir. 2018); United States v. Foust, 989 F.3d 842, 846 (10th Cir.), *cert. denied*, 142 S. Ct. 294 (2021).

223. 2020 NIST Report, *supra* note 110, at 8.

> and arriving at an opinion regarding the writership of the questioned writing.[224]

Examiners believe that the various features they examine, such as letter formation and size, and the inter- and intra-word spacing, are more variable in the writing of different individuals than in the writing of the same individual, but today there is widespread acknowledgment that the statistical properties of such variabilities have not been rigorously studied.[225] There is no universally accepted number of points of similarity required to conclude the writings came from the same source. Additionally, although there is an expected range of variability for a single writer, there are no specific standards or procedures for determining that range.[226]

After evaluation, the examiner "expresses an opinion indicating [their] subjective confidence in the process outcome."[227] There are five opinions possible, according to the 2020 NIST Report: (1) Identification; (2) Probably did write; (3) Inconclusive; (4) Probably did not write; and (5) Elimination.[228] In contrast, the Scientific Working Group for Forensic Document Examination (SWGDOC) gives examiners nine options that may be expressed, along with associated descriptions.[229] For its part, the FBI laboratory uses five categories that "collapse" the nine possible opinions on the SWGDOC scale.[230] As of 2021, DOJ requires examiners to offer any of the following five conclusions: (1) Source Identification (i.e., identified); (2) Support for a Common Source; (3) Inconclusive; (4) Support for Different Sources; and (5) Source Exclusion (i.e., excluded).[231] DOJ has also limited the certainty that experts may use, consistent with other forensic feature comparison methods.[232] In short, just as there is no consensus about the number of specific points of similarity to determine authorship, there is no nationwide consensus about the range of possible opinions.

224. *Id.* at 9.

225. *Id.* at 14.

226. *Id.*

227. *Id.* at 24.

228. *Id.*

229. *See, e.g.*, 2009 NRC Report, *supra* note 37, at 166; 2020 NIST Report, *supra* note 110, at 25–26, listing: 1. Identification; 2. Strong probability; 3. Probable; 4. Indications; 5. No conclusion; 6. Indications did not; 7. Probably did not; 8. Strong probability did not; and 9. Elimination.

230. 2020 NIST Report, *supra* note 110, at 25, citing SWGDOC, Version 2013-2.

231. U.S. Dep't of Justice, Uniform Language for Testimony and Reports for Testimony and Reports for Forensic Document Examination 2 (2021), https://perma.cc/KMC4-6KG3.

232. *Id.* at 4 ("Therefore, an examiner shall not: assert that a 'source identification' or a 'source exclusion' conclusion is based on the 'uniqueness' of an item of evidence; use the terms 'individualize' or 'individualization' when describing a source conclusion; assert that two or more bodies of writing were prepared by the same writer to the exclusion of all other writers.").

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of handwriting analysis, given that these will frame the litigation judges will face over admission of FDE testimony.

As is the case with other feature comparison methods, critiques of the foundational reliability of handwriting analysis have focused on its relative subjectivity, lack of empirical data underlying assumptions about rarity of characteristics, potential for contextual bias to affect examinations, and lack of accurate error rate estimates. The 2009 NRC Report concluded that the technique was useful but not yet scientifically valid for determining authorship:

> The scientific basis for handwriting comparisons needs to be strengthened. Recent studies have increased our understanding of the individuality and consistency of handwriting . . . and suggest that there may be a scientific basis for handwriting comparison, at least in the absence of intentional obfuscation or forgery. Although there has been only limited research to quantify the reliability and replicability of the practices used by trained document examiners, the committee agrees that there may be some value in handwriting analysis.[233]

A long-time forensic document examiner himself agreed with this assessment, writing in a 2018 article that the method could be very reliable under the right conditions and with more research but that "forensic handwriting analysis still lacks robust, ground truth studies that provide empirical support for the reliability of many of the tasks routinely performed."[234]

The latest governmental evaluation of handwriting analysis, the 2020 NIST Report, focused its foundational reliability concerns on the relative subjectivity of the method, the resulting potential for contextual bias, and the need for empirical testing to determine the extent of both intra- and inter-examiner variability (that is, repeatability and reproducibility, two aspects of what is often called reliability, as compared to accuracy).[235] To provide trustworthy estimates of repeatability and reproducibility, studies should "compare the performance within and between FDEs in their judgments on the same samples of handwriting against ground truth."[236] The report urges that multiple independent laboratories work together on these problems by using the same methods and materials.[237] The NIST report ultimately urges that both "black box studies" (to calculate an accurate error rate estimate) and "white box studies" (to understand the steps examiners are taking, with an eye toward developing standards that then must be applied when shown to lead to accurate results) are needed, and

233. 2009 NRC Report, *supra* note 37, at 166–67.
234. Sulner, *supra* note 221, at 714.
235. 2020 NIST Report, *supra* note 110, at 52.
236. *Id.* at 53.
237. *Id.*

at least one has been performed since 2020 (when the NIST Human Factors report was released).[238]

The 2020 NIST Report also expressed concern over the method's reliability as applied, and in particular, the use by examiners of language that suggests source attribution or that otherwise goes beyond the scientifically supportable bounds of the method or overstates the evidence. The report cautioned, for example, that the expressions "uniqueness" and "individualization" do not have a settled meaning in forensic science, and their casual use can lead to misunderstandings in court and an "exaggeration of the strength of such evidence."[239] The report also concluded that "empirical research and statistical reasoning do not support source attribution to the exclusion of all others." Instead it forcefully recommended that FDEs "must not report or testify, directly or by implication, that questioned handwriting has been written by an individual (to the exclusion of all others)."[240] It likewise recommended that language about uniqueness and individualization should give way to a critical evaluation of the "rarity of the features," on a continuum of rare to common. "A more contemporary view . . . individuality is defined with respect to the probability of observing writing profiles of two individuals that are indistinguishable using the specified comparison method."[241] Additionally, there is movement away from the conventional approach to handwriting analysis and toward a more empirical, neurobiological approach that would permit hypothesis generation and testing of relevant principles.[242]

Finally, with respect to reliability as applied, the 2020 NIST Report dedicated a full section to the several types of cognitive bias that could arise in FDE because of its layers of relatively high subjectivity.[243] The report cautioned that such bias is a "legitimate cause for concern" that must be addressed by forensic

238. *See, e.g.*, R. Austin Hicklin et al., *Accuracy and Reliability of Forensic Handwriting Comparisons*, 119 Proc. Nat'l Acad. Scis. e2119944119 (2022), https://doi.org/10.1073/pnas.2119944119.

239. 2020 NIST Report, *supra* note 110, at 47.

240. *Id.* DOJ's ULTR for Forensic Document Examination continues to permit an examiner to offer a conclusion on "source identification" but must not assert that "two or more bodies of writing were prepared by the same writer to the exclusion of all other writers." U.S. Dep't of Justice, Uniform Language for Testimony and Reports for Testimony and Reports for Forensic Document Examination 3 (2021), https://perma.cc/KMC4-6KG3.

241. 2020 NIST Report, *supra* note 110, at 46 (citing Sargur Srihari et al., *Individuality of Handwriting*, 47 J. Forensic Scis. 856 (2002)).

242. *See, e.g.*, Michael P. Caligiuri & Linton A. Mohammed, The Neuroscience of Handwriting 35–57 (2012).

243. 2020 NIST Report, *supra* note 110, at 30. The NIST report dedicates an entire section (2.1) to the multiple types of contextual bias in FDE. *Id.* at 30–44. *See also* D. Michael Risinger et al., *The* Daubert/Kumho *Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 Calif. L. Rev. 1 (2002), https://doi.org/10.2307/3481305; Adele Quigley-McBride et al., *A Practical Tool for Information Management in Forensic Decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in Casework*, 4 Forensic Sci. Int'l: Synergy 100216 (2022), https://doi.org/10.1016/j.fsisyn.2022.100216.

laboratories through standards, protocols like the "contextual information management" designed to control information flow to and among examiners, and documentation of steps taken to avoid such bias.[244] The report made several recommendations along these lines and urged research to study the problem.[245] There are different suggested approaches to combat these problems—notably Dr. Itiel Dror's hierarchy of expert performance method (HEP).[246]

Another legal question that has arisen is whether FDE expert testimony is "helpful to the jury" for Rule 702 purposes. Forensic handwriting analysis is a unique discipline because handwriting is a fundamental aspect of everyday life, and thus, ordinary people are capable of comparing some aspects of handwriting samples.[247] Laypersons are usually capable of identifying gross characteristics, but forensic handwriting analysis focuses on nuanced details that laypersons are often unable to correctly identify.[248] A 2022 FBI-funded study found that adequately trained forensic handwriting experts are more accurate than those with minimal training and that examiners with less than two years of formal training had higher error rates and were more likely to make definitive, unqualified conclusions compared to those with at least two years of formal training.[249] A 2018 study reached a similar conclusion when comparing novice and expert examiners, but found that the overall error rate among expert examiners was still high enough to question the trustworthiness of handwriting comparison evidence in the courts.[250]

244. 2020 NIST Report, *supra* note 110, at 34.

245. *Id.* at 35–36. For a detailed discussion of how the contextual information management (CIM) should work in FDE, see *id.* at 36–44.

246. *See, e.g.*, Itiel Dror, *A Hierarchy of Expert Performance*, 5 J. Applied Res. Memory & Cognition 121 (2016), https://doi.org/10.1016/j.jarmac.2016.03.001 (creating a hierarchy of expert performance to identify weaknesses in an individual's performance and compare experts across domains).

247. 2020 NIST Report, *supra* note 110, at viii.

248. *See, e.g.*, State v. Cooke, 914 A.2d 1078, 1101 (Del. Super. Ct. 2007) ("Handwriting comparison is beyond a lay person's general knowledge."); Diane Harrison, Ted M. Burkes & Danielle P. Sieger, *Handwriting Examination: Meeting the Challenges of Science and the Law*, Forensic Sci. Commc'ns (Oct. 2009), https://perma.cc/J2Z2-W7D5.

249. R. Austin Hicklin et al., *Accuracy and Reliability of Forensic Handwriting Comparisons*, 119 Proc. Nat'l Acad. Scis. e2119944119 (2022), https://doi.org/10.1073/pnas.2119944119.

250. Kristy Martire et al., *What Do the Experts Know? Calibration, Precision, and the Wisdom of Crowds Among Forensic Handwriting Experts*, 25 Psych. Bull. Rev. 2346, 2353 (2018), https://doi.org/10.3758/s13423-018-1448-3:

> On the one hand, there is some evidence that handwriting experts will be able to estimate the frequency of occurrence for handwriting features better than novices. However, even the single best performing participant produced an average deviation of 18.5% from the true value. On the other hand, this number is considerably lower than would be expected by chance (25%).

Some examiners incorporate handwriting pattern recognition technology to assist in their analysis.[251] This technology can outperform laypersons[252] and can be as accurate as an expert examiner, depending on the complexity of the task.[253] These systems will only become more advanced as machine learning algorithms enhance their accuracy and reliability.[254] Another area of research has been computer-driven forensics linguistics to identify authors.[255] For instance, a Swiss company, OrphAnalytics SA, recently released a software package for conducting stylometric analysis to support the attribution of text to a particular author. While promising, these techniques have yet to be used forensically, and as with other forms of artificial intelligence, they too will have shortcomings. The 2020 NIST Report calls for FDEs to integrate these tools into their casework and "collaborate with the computer science and engineering communities to develop and validate applicable, user-friendly, automated systems."[256] Training programs requiring proficiency in the use of these systems, combined with interdisciplinary research and development, should provide courts with additional metrics to evaluate accuracy and reliability of handwriting comparisons in the future.

Case Law Development

Although nineteenth-century courts were skeptical of handwriting expertise,[257] the twentieth century saw a marked shift, as testimony in leading cases like the Lindbergh kidnapping helped the discipline gain judicial acceptance.[258] There was little dispute that handwriting testimony was admissible at the time the

251. 2020 NIST Report, *supra* note 110, at 68–71.

252. Harrison et al., *supra* note 248 (citing Sargur Srihari et al., *On the Discriminability of the Handwriting of Twins*, 53 J. Forensic Scis. 430 (2008), https://doi.org/10.1111/j.1556-4029.2008.00682.x).

253. 2020 NIST Report, *supra* note 110, at 70–71 (citing Muhammad Imran Malik et al., *Man vs. Machine: A Comparative Analysis for Signature Verification*, 24 J. Forensic Document Examination 21 (2004)).

254. 2020 NIST Report, *supra* note 110, at 71.

255. *See, e.g.*, Janet Ainsworth & Patrick Juola, *Who Wrote This?: Modern Forensic Authorship Analysis as a Model for Valid Forensic Science,* 96 Wash. U. L. Rev. 1159, 1169–72 (2019) (language has many "objectively identifiable features"); Patrick Juola, *Verifying Authorship for Forensic Purposes: A Computational Protocol and its Validation*, 325 Forensic Sci. Int'l 110824 (2021), https://doi.org/10.1016/j.forsciint.2021.110824 (claiming a measured accuracy of 77% across more than 32,000 different document pairs); Cami Fuglsby et al., *Elucidating the Relationships Between Two Automated Handwriting Feature Quantification Systems for Multiple Pairwise Comparisons,* 67 J. Forensic Scis. 642 (2022), https://doi.org/10.1111/1556-4029.14914.

256. 2020 NIST Report, *supra* note 110, at 72.

257. *See* Strother v. Lucas, 31 U.S. 763, 767 (1832); Phoenix Fire Ins. Co. v. Philip, 13 Wend. 81, 82–84 (N.Y. Sup. Ct. 1834).

258. *See* Risinger et al., *Exorcism of Ignorance, supra* note 210, at 771.

Federal Rules of Evidence were enacted in 1975. Rule 901(b)(3) recognized that a document could be authenticated by an expert, and the drafters explicitly mentioned handwriting comparison through the "testimony of expert witnesses."[259]

The first significant admissibility challenge under *Daubert* was in *United States v. Starzecpyzel*.[260] In that case, the district court concluded that "forensic document examination, despite the existence of a certification program, professional journals and other trappings of science, cannot, after *Daubert*, be regarded as 'scientific . . . knowledge.'"[261] Nevertheless, the court did not exclude handwriting comparison testimony. Instead, the court admitted the individuation testimony as nonscientific "technical" evidence.[262]

Starzecpyzel prompted more litigation that questioned the lack of empirical validation in the field.[263] For many years, there was a three-way split of authority. The majority of courts permitted examiners to express individuation opinions.[264] As one court noted, "all six circuits that have addressed the admissibility of handwriting expert [testimony] . . . [have] determined that it can satisfy the reliability threshold" for nonscientific expertise.[265] In 2021, the Tenth Circuit, in *United States v. Foust*, held that handwriting comparison was properly admitted, relying largely on the "widespread acceptance of handwriting comparison through the years" and citing the importance of the general acceptance factor to its decision.[266] The *Foust* opinion explained that the proposed testimony fell short in the "standards, peer review, and error rate" factors, but upheld the trial

259. Fed. R. Evid. 901(b)(3) advisory committee's note.

260. 880 F. Supp. 1027 (S.D.N.Y. 1995).

261. *Id.* at 1038.

262. *Id.* at 1047.

263. *See, e.g.*, United States v. Hidalgo, 229 F. Supp. 2d 961, 967 (D. Ariz. 2002):

> Because the principle of uniqueness is without empirical support, we conclude that a document examiner will not be permitted to testify that the maker of a known document is the maker of the questioned document. Nor will a document examiner be able to testify as to identity in terms of probabilities.

264. *See, e.g.*, United States v. Prime, 363 F.3d 1028, 1033 (9th Cir. 2004); United States v. Crisp, 324 F.3d 261, 265–71 (4th Cir. 2003); United States v. Jolivet, 224 F.3d 902, 906 (8th Cir. 2000) (affirming the introduction of expert testimony that it was likely that the accused wrote the questioned documents); United States v. Velasquez, 64 F.3d 844, 848–52 (3d Cir. 1995); United States v. Ruth, 42 M.J. 730, 732 (A. Ct. Crim. App. 1995), *aff'd on other grounds,* 46 M.J. 1 (C.A.A.F. 1997); United States v. Morris, No. 06-87-DCR, 2006 WL 2054585, *2 (E.D. Ky. July 20, 2006); Orix Fin. Servs. v. Thunder Ridge Energy, Inc., No. 01 Civ. 4788, 2006 WL 587483 (S.D.N.Y. Mar. 8, 2006).

265. *Prime*, 363 F.3d at 1034.

266. United States v. Foust, 989 F.3d 842, 846 (10th Cir.), *cert denied*, 142 S. Ct. 294 (2021). For more about courts' reliance on a "long history of use" as justification to admit forensic evidence, see Jane Campbell Moriarty, *Deceptively Simple: Framing, Intuition, and Judicial Gatekeeping of Forensic Feature-Comparison Methods Evidence,* 86 Fordham L. Rev. 1687 (2018).

judge's decision to admit the evidence. Explaining that the *Daubert* factors were meant to be helpful and not definitive, the Tenth Circuit held that the trial court did not abuse its discretion in admitting the evidence. Echoing *Starzecpyzel,* the court noted "handwriting comparison is not a traditional science."[267] So, too, the Sixth Circuit in *United States v. Mallory* held that handwriting comparison was properly admitted, as it was based on "knowledge and experience to answer the extremely practical question of whether a signature is genuine or forged."[268] The court reasoned that since experts "see things in handwriting that laypeople do not—both because of analysts' training in the minutiae of loops, swoops, and dotted 'i's, and because of the volume of handwriting they inspect," such testimony is helpful to the jury. Like the Tenth Circuit in *Foust*, the Sixth Circuit recognized that "handwriting analysis may not boast the 'empirical' support underpinning scientific disciplines," but ruled that it was properly admitted as a proper area of expertise based on specialized or technical expertise.[269]

Both *Foust* and *Mallory* downplayed the concerns raised about foundational reliability, choosing to cite older, pre-NRC and PCAST report cases, and to cast the evidence as technical or specialized rather than scientific. Both courts determined that as the specialty was not scientific, the traditionally applied *Daubert* factors were of less concern.[270] In light of the 2023 amendments to Federal Rule of Evidence 702, as discussed earlier, courts may be required to more robustly address the concerns about foundational reliability that this specialty poses. These two cases also highlight potential questions courts may have about the need for proof of reliability as applied. Both courts appeared to rely on the fact that the experts were qualified and had years of experience. Indeed, many other courts have relied on experience and qualifications as proxies for reliability of FDE as applied.[271] There are two potential concerns with this approach. First, some commentators have argued that experience alone, at least in relatively subjective disciplines, is not a substitute for proficiency testing in establishing expert qualification. Second, the fact that an expert is proficient in a method does not establish that the method was reliably applied in a given case.[272] Reliability as applied could instead be established through internal validation, case file documentation, standards and protocols showing avoidance of contextual bias, the

267. *Foust*, 989 F.3d at 846–47.

268. United States v. Mallory, 902 F.3d 584, 593 (6th Cir. 2018).

269. *Id.*

270. *Id.* at 593–94.

271. *See also* Giannelli et al., Scientific Evidence, *supra* note 209, § 1.03[2] (noting the tendency of some courts to elide qualifications with reliability inquiries, particularly with experiential-knowledge-based specialties).

272. Garrett & Mitchell, *supra* note 19, at 909 ("Experts should be qualified based on empirical evidence of their proficiency before addressing whether their methods used and conclusions reached are valid and reliable.").

nature of the samples examined, and the examiner's own testimony as to what they did and the conclusions they drew.

Other courts have been more amenable to hearing challenges to FDE. Although the excluded testimony at issue in *United States v. Johnsted* related to hand printing rather than hand writing, the court expressed doubts about the foundational reliability of the entire field of handwriting as well:

> This lack of testing has serious repercussions on a practical level: because the entire premise of interpersonal individuality and intrapersonal variations of handwriting remains untested in reliable, double blind studies, the task of distinguishing a minor intrapersonal variation from a significant interpersonal difference—which is necessary for making an identification or exclusion—cannot be said to rest on scientifically valid principles. The lack of testing also calls into question the reliability of analysts' highly discretionary decisions as to whether some aspect of a questioned writing constitutes a difference or merely a variation; without any proof indicating that the distinction between the two is valid, those decisions do not appear based on a reliable methodology. With its underlying principles at best half-tested, handwriting analysis itself would appear to rest on a shaky foundation.[273]

In another federal district court case, *Almeciga v. Center for Investigative Reporting, Inc.*,[274] the court excluded on *Daubert* grounds testimony about whether writing was real or disguised, opining that the field is relatively subjective and lacks adequate testing:

> [T]here are no studies, to this Court's knowledge, that have evaluated the extent to which the angle at which one writes or the curvature of one's loops distinguish one person's handwriting from the next. Precisely what degree of variation falls within or outside an expected range of natural variation in one's handwriting—such that an examiner could distinguish in an objective way between variations that indicate different authorship and variations that do not—appears to be completely unknown and untested. Ditto the extent to which such a range is affected by the use of different writing instruments or the intentional disguise of one's natural hand or the passage of time. Such things could be tested and studied, but they have not been; and this by itself renders the field unscientific in nature.[275]

273. United States v. Johnsted, 30 F. Supp. 3d 814, 818 (W.D. Wis. 2013). *See also* United States v. Fujii, 152 F. Supp. 2d 939, 940 (N.D. Ill. 2000) (holding expert testimony concerning Japanese handprinting inadmissible; "Handwriting analysis does not stand up well under the *Daubert* standards. Despite its long history of use and acceptance, validation studies supporting its reliability are few, and the few that exist have been criticized for methodological flaws."); United States v. Lewis, 220 F. Supp. 2d 548 (S.D. W. Va. 2002); United States v. Saelee, 162 F. Supp. 2d 1097 (D. Alaska 2001); Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401 (S.D.N.Y. 2016).

274. 185 F. Supp. 3d 401 (S.D.N.Y. 2016).

275. *Id.* at 419.

The court also relied on what it described as a dearth of peer review and added that existing studies are "not sufficiently robust or objective to lend credence to the proposition that handwriting comparison is a scientific discipline."[276] The court stated that the known error rate is poor and even worse when there is a question of a disguised signature.[277] Finally, the court discussed the lack of standards as to how many similarities are needed for a "match" or how many exemplars must be considered. In the court's view, many conclusions by questioned document examiners are highly subjective, "bottom line" decisions whether there is a "match." Summing up, the court wrote: "It remains the case that the methodology has not been subject to adequate testing or peer review, that error rates for the task at hand are unacceptably high, and that the field sorely lacks internal controls and standards, and so forth."[278] The court did not rule categorically that all handwriting comparison is inadmissible but noted the importance for caution in admitting testimony from an FDE, particularly on an opinion on authorship. More specifically, the court suggested that a trial judge "should not [admit questioned document testimony] without carefully evaluating whether the examiner has actual expertise in regard to the specific task at hand," including proficiency testing.[279] The court concluded that in this case, the evidence of the witness's expertise was insufficient.[280]

Some district courts have endorsed a third view. These courts limit the reach of the examiner's opinion, permitting expert testimony about similarities and dissimilarities between exemplars but not an ultimate conclusion that the defendant was the author (common authorship opinion) of the questioned document.[281] The expert is allowed to testify about "the specific similarities and

276. *Id.* at 420–21.

277. *Id.* at 422. Studies "suggest that while forensic document examiners might have some arguable expertise in distinguishing an authentic signature from a close forgery, they do not appear to have much, if any, facility for associating an author's natural handwriting with his or her disguised handwriting."

278. *Id.* at 424. *Accord Johnsted*, 30 F. Supp. 3d 814.

279. *Almeciga*, 185 F. Supp. 3d at 424.

280. *But see* United States v. Pitts, No. 16-CR-550, 2018 WL 1116550 (E.D.N.Y. Feb. 26, 2018) (noting that many courts have admitted such evidence in the Second Circuit and distinguishing *Almeciga* as it involved a potentially forged signature, not just a comparison of known and unknown samples).

281. *See, e.g.*, United States v. Oskowitz, 294 F. Supp. 2d 379, 384 (E.D.N.Y. 2003) ("Many other district courts have similarly permitted a handwriting expert to analyze a writing sample for the jury without permitting the expert to offer an opinion on the ultimate question of authorship."); United States v. Rutherford, 104 F. Supp. 2d 1190, 1194 (D. Neb. 2000):

> [T]he Court concludes that [the examiner]'s testimony meets the requirements of Rule 702 to the extent that he limits his testimony to identifying and explaining the similarities and dissimilarities between the known exemplars and the questioned documents. [The examiner] is precluded from rendering any ultimate conclusions on authorship of the questioned documents and is similarly precluded from testifying to the degree of confidence or certainty on which his opinions are based.

idiosyncrasies between the known writings and the questioned writings, as well as testimony regarding, for example, how frequently or infrequently in his experience, [the expert] has seen a particular idiosyncrasy."[282] As the justification for this limitation, these courts often state that the examiners' claimed ability to individuate lacks "empirical support."[283]

As explained at the end of the section titled "Fingerprint Evidence" above, the 2023 amendment to Rule 702 may require courts to engage in more robust gatekeeping to evaluate both the foundational reliability and the reliability as applied of handwriting comparison evidence. The new amendment to Rule 702 may require a reconsideration of prior decisions focusing solely on weight and not admissibility of such evidence. As the advisory committee's notes clarify, "many courts have held that the critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a)."[284] As the notes recommend, "[j]udicial gatekeeping is essential." As explained earlier, several gatekeeping options are available to the courts, along with the option of providing jury instructions to aid jurors in evaluating the evidence. With the enactment of the 2023 amendment to Rule 702, the judge must be satisfied that the proponent of the evidence has met each of the rule's requirements by a preponderance. If the judge is not satisfied, she should exclude those opinions. Additionally, the judge may appropriately limit any opinion to exclude overstatement about the conclusions.

Firearms and Toolmark Analysis

Introduction

When ammunition travels through a firearm, the firearm leaves indentations or marks on the ammunition that arguably may be used to link it with a particular firearm or class of firearms. To that end, firearm and toolmark (sometimes referred to as FATM or FA/TM) examiners analyze fired ("spent") ammunition found at crime scenes. They compare the marks on such ammunition either to the marks on other ammunition, to determine if both may have been fired from the same firearm or same category of firearm, or to a test fire from a particular suspected weapon, to determine if the ammunition might have been fired from

See also United States v. Hines, 55 F. Supp. 2d 62, 69 (D. Mass. 1999) (expert testimony concerning the general similarities and differences between a defendant's handwriting exemplar and a stick-up note was admissible while the specific conclusion that the defendant was the author was not).

282. United States v. Van Wyk, 83 F. Supp. 2d 515, 524 (D.N.J. 2000).

283. United States v. Hidalgo, 229 F. Supp. 2d 961, 967 (D. Ariz. 2002).

284. Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

that weapon or a similar class of weapon. A related technique is used to compare the marks left on a surface of interest (like at a crime scene) to the marks left by a known tool (like a screwdriver) to determine if the known tool might have left the mark in question.

For decades, firearms and toolmark identification evidence has been widely accepted by courts.[285] As with other feature comparison evidence, it was accepted with little critical scrutiny of its reliability and accuracy. More recently, the technique has been increasingly challenged. As with many disciplines, it is least controversial when used for excluding firearms, rather than identifying an individual firearm—exclusion as opposed to inclusion.

Concerns about the accuracy of firearm/toolmark analysis when used for source attribution are not merely hypothetical. The National Registry of Exonerations documents wrongful convictions based on firearms and toolmark analysis.[286] Most notable is likely the wrongful conviction of Anthony Ray Hinton, who spent thirty years on death row for a murder he did not commit based on a firearm misidentification, before the U.S. Supreme Court reversed his conviction.[287] Firearms misidentifications have also resulted in crime lab audits and lost accreditations.[288]

285. *See* Gardner v. United States, 140 A.3d 1172, 1183 (D.C. 2016) (discussing how District of Columbia courts have "allowed the admission of expert testimony concerning ballistics comparison matching techniques" for "decades") (citing Laney v. United States, 294 F. 412, 416 (D.C. Cir. 1923)).

286. *See, e.g.*, National Registry of Exonerations, *Patrick Pursley*, https://perma.cc/Q4VG-E5JC (last updated 4/29/2023); People v. Pursley, 2018 IL App (2d) 170227-U (2018). *See also* Brandon Garrett, *Siggers' Firearms Exoneration*, Duke L. Forensic Forum (Oct. 23, 2018), https://perma.cc/2HHS-HUYC (wrongful conviction of Darrell Siggers); Craig Cooley & Gabriel Oberfield, *Increasing Forensic Evidence's Reliability and Minimizing Wrongful Convictions: Applying* Daubert *Isn't the Only Problem*, 43 Tulsa L. Rev. 285, 337–38 (2007) (wrongful conviction of Charles Stielow); Simon A. Cole, *Implicit Testing: Can Casework Validate Forensic Techniques?*, 46 Jurimetrics J. 117, 126–27 (2006). The Association of Firearm and Tool Mark Examiners (AFTE) has their own journal to publish studies; for concessions from the AFTE journal, see Bruce Moran, *A Report on the AFTE Theory of Identification & Range of Conclusions for Tool Mark Identification & Resulting Approaches to Casework*, 34 AFTE J. 227 (2002) ("In the 1980s some striated toolmark misidentifications resulting from a poor understanding of toolmark criteria for identification were experienced. An increasing need to address problems of applying subjective criteria became apparent."); Evan E. Hodge, *Guarding Against Error*, 20 AFTE J. 290 (1988) (noting that "most of us [firearms examiners] know someone who has committed serious error" and describing misidentification by another examiner of the wrong .45 caliber firearm because of cognitive bias and pressure from prosecutors); Lowell Bradford, *Forensic Firearms Identification: Competence or Incompetence*, 11(2) AFTE J. 12 (1979) ("[a]n appalling number of misidentifications have been found in the firearm identification field").

287. *See* Hinton v. Alabama, 571 U.S. 263 (2014); National Registry of Exonerations, *Anthony Ray Hinton*, https://perma.cc/YJ3J-UQRU (last updated Aug. 1, 2017).

288. In 2008, the Michigan State Police Forensic Science Division conducted an audit of the Detroit Police Department's firearms unit at the request of the Wayne County Prosecutor's Office and the Detroit Police Department chief. The audit included a random reanalysis of 283 cases. In 10% of the reanalyzed cases, the firearms unit had made false identifications or false exclusions.

Basic terms. Typically, three types of firearms—rifles, handguns, and shotguns—are encountered in criminal investigations. Rifles and handguns are classified according to their caliber. The caliber is the diameter of the bore of the firearm; the caliber is expressed in either hundredths or thousandths of an inch (e.g., .22, .45, .357 caliber) or millimeters (e.g., 7.62 mm). The two major types of handguns are revolvers[289] and semiautomatic pistols. A major difference between the two is that when a semiautomatic pistol is fired, the cartridge case is automatically ejected and, if recovered at the crime scene, could help link the case to the firearm from which it was fired. In contrast, when a revolver is discharged the case is not ejected. In terms of ammunition, rifle and handgun cartridges consist of the projectile (bullet),[290] case,[291] propellant (powder), and primer. The primer contains a small amount of an explosive mixture, which detonates when struck by the firing pin. When the firing pin detonates the primer, an explosion occurs that ignites the propellant. The most common modern propellant is smokeless powder.

The barrels of modern rifles and handguns are rifled; that is, parallel spiral grooves are cut into the inner surface, or bore, of the barrel. The surfaces between the grooves are called lands. The lands and grooves twist in a direction: right twist or left twist. For each type of firearm produced, the manufacturer specifies the number of lands and grooves, the direction of twist, the angle of

Michigan State Police Forensic Science Division, *Audit of the Detroit Police Department Forensic Services Laboratory Firearms Unit* (2008); *see also* Nick Bunkley, *Detroit Police Lab is Closed After Audit Finds Serious Errors in Many Cases*, N.Y. Times, Sept. 25, 2008, https://www.nytimes.com/2008/09/26/us/26detroit.html. Similarly, a failed proficiency test by a firearms examiner in 2017 launched multiple audits of the D.C. Department of Forensic Sciences. The audits revealed that three firearms examiners had misidentified cartridge cases in ongoing cases. The D.C. Department of Forensic Sciences lost its accreditation in 2021, after which the laboratory fired all of its firearms personnel, and set to review every firearms case from the last decade. *See* Spencer S. Hsu & Keith L. Alexander, *Forensic Errors Trigger Reviews of D.C. Crime Lab Ballistics Unit, Prosecutors Say*, Wash. Post, Mar. 24, 2017, https://perma.cc/28FH-LWTT; Jack Moore, *Sweeping Report Urges DC to Review Every Case Handled by Firearms, Fingerprint Units at Troubled Crime Lab*, WTOP News, Dec. 14, 2021, https://perma.cc/GFN8-BYQK; Jack Moore, *Officials Now Expect DC Crime Lab to Remain Sidelined Until Next Spring*, WTOP News, Mar. 31, 2022, https://perma.cc/6LMQ-P8RM.

289. Revolvers have a cylindrical magazine that rotates behind the barrel. The cylinder typically holds five to nine cartridges, each within a separate chamber. When a revolver is fired, the cylinder rotates and the next chamber is aligned with the barrel. A single-action revolver requires the manual cocking of the hammer; in a double-action revolver the trigger cocks the hammer. *See generally* BulletPoints Project, *Guns 101*, https://perma.cc/6A43-DYDT (last visited Dec. 5, 2024) (state-funded firearms injury prevention site with basic information about firearms).

290. Bullets are generally composed of lead and small amounts of other elements (hardeners). They may be completely covered (jacketed) with another metal or partially covered (semi-jacketed). *See id*. (discussing types and basics of ammunition).

291. Cartridge cases are generally made of brass. *Id*. They are sometimes referred to as "casings" by courts and laypeople, but firearms examiners do not use that term.

twist (pitch), the depth of the grooves, and the width of the lands and grooves. As a bullet passes through the bore, the lands and grooves force the bullet to rotate, giving it stability in flight and thus increased accuracy. Shotguns are "smooth-bore" firearms; that is, they do not have lands and grooves.

The Firearms Comparison Method and Its Claims

Bullet identification involves a comparison of the evidence bullet and a test bullet fired from the firearm.[292] The two bullets are examined by means of a comparison microscope, which permits a split-screen view of the two bullets and manipulation in order to attempt to align the striations (marks) on the two bullets.

The theory behind this comparative analysis is that barrels are machined during the manufacturing process, and imperfections in the tools used in the machining process are imprinted on the bore. The subsequent use of the firearm could add further individual imperfections. For example, mechanical action (erosion) caused by the friction of bullets passing through the bore of the firearm could produce accidental imperfections. Similarly, chemical action (corrosion) caused by moisture (rust), as well as primer and propellant chemicals, could produce other imperfections. However, the prevalence of certain imperfections remains unknown.

When a bullet is fired, microscopic striations are imprinted on the bullet surface as it passes through the bore of the firearm. These bullet markings are produced by the imperfections in the bore. Because these imperfections are randomly produced, examiners assume that they are unique to each firearm. Currently, there is no statistical basis for this assumption.

The condition of a firearm or evidence bullet may preclude a conclusion. For example, there may be insufficient marks on the bullet or, because of mutilation, an insufficient amount of the bullet may have been recovered. Likewise, if the bore of the firearm has changed significantly as a result of erosion or corrosion, a conclusion may be impossible. (Unlike fingerprints, firearms change over time.) In these situations, the examiner may render a "no conclusion" determination.

Firearms comparisons are made based on examinations of so-called class, subclass, and individual characteristics of bullet or cartridge cases. The class characteristics of a firearm result from design factors prior to manufacture. They include caliber and rifling specifications: (1) the land and groove diameters,

292. Test bullets are obtained by firing a firearm into a recovery box or bullet trap, which is usually filled with cotton, or into a recovery tank, which is filled with water. *See generally* NIST, Forensic Science Program, *Firearms and Toolmarks*, https://perma.cc/9ATM-CZNX (last updated Dec. 1, 2024) (discussing basics of firearms testing).

(2) the direction of rifling (left or right twist), (3) the number of lands and grooves, (4) the width of the lands and grooves, and (5) the degree of the rifling twist. Generally, a .38-caliber bullet with six land and groove impressions and with a right twist could have been fired only from a firearm with these characteristics. Such a bullet could not have been fired from a .32-caliber firearm, or from a .38-caliber firearm with a different number of lands and grooves or a left twist. According to the Association of Firearm and Tool Mark Examiners (AFTE),[293] subclass characteristics are discernible surface features that are more restrictive than class characteristics in that they are (1) "produced incidental to manufacture," (2) "relate to a smaller group source (a subset to which they belong)," and (3) can arise from a source that changes over time.[294] The AFTE states that "[c]aution should be exercised in distinguishing subclass characteristics from class characteristics."[295] Finally, according to the AFTE, imperfections in the machining process and subsequent use of a firearm can cause individual characteristics that are thought to be unique to a particular firearm. Experts have warned that "caution should be exercised in distinguishing subclass characteristics from individual characteristics."[296]

In terms of examiners' ultimate opinions, the FBI's ULTR on firearms and toolmark examinations directs examiners to report one of three conclusions: (1) source identification; (2) source exclusion; or (3) inconclusive.[297] AFTE, in contrast, advises examiners that a source conclusion can be made when "sufficient agreement" exists in the patterns of two sets of marks. Agreement is sufficient when it "exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool," such that "the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility."[298]

Although a conclusion about what marks exist is in one sense based on relatively objective data (the striations on the bullet surface), the AFTE explains that

293. AFTE is the leading professional organization in the field. There is also the Organization of Scientific Areas (OSAC) Firearms and Toolmarks Subcommittee, which promulgates standards and guidelines for examiners.

294. *Theory of Identification, Association of Firearm and Toolmark Examiners*, 30 AFTE J. 86, 88 (1998).

295. *Id*.

296. Robert M. Thompson, Program Manager, Forensic Data Systems, Office of Law Enforcement Standards, NIST, *Firearm Identification in the Forensic Laboratory*, Nat'l Dist. Att'ys Ass'n (2010), https://perma.cc/P37Y-ASNM.

297. *See* U.S. Dep't of Justice, Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Pattern Examination (2023), https://perma.cc/4N57-TDAG.

298. Association of Firearm & Tool Mark Examiners, *AFTE Theory of Identification*, https://perma.cc/VTS2-HER3 (last visited Dec. 5, 2024).

the examiner's ultimate conclusion as to whether a projectile was fired from a particular weapon is a largely subjective judgment. The AFTE describes the traditional pattern recognition methodology as "subjective in nature, founded on scientific principles and based on the examiner's training and experience."[299] There are no objective criteria governing this determination: "Ultimately, unless other issues are involved, it remains for the examiner to determine for himself the modicum of proof necessary to arrive at a definitive opinion."[300]

Another firearms comparison technique is cartridge case identification. Cartridge case identification is based on the same theory of random markings as bullet identification.[301] As with barrels, the belief is that defects produced in manufacturing leave distinctive characteristics on the breech face, firing pin, chamber, extractor, and ejector. Later use of the firearm is believed to produce additional defects. When the trigger is pulled, the firing pin strikes the primer of the cartridge, causing the primer to detonate. This detonation ignites the propellant (powder). In the process of combustion, the powder is converted rapidly into gases. The pressure produced by this process propels the bullet from the weapon and also forces the "base" (or "head") of the cartridge case backward against the breech face, imprinting breech face marks on the base of the cartridge case.[302] In turn, cartridge case identification involves a comparison of the case recovered at the crime scene and a test case obtained from the firearm after it has been fired. Shotgun shell cases may be analyzed in this way, as well. As in bullet identification, the comparison microscope is used in the examination, and the examiner's findings are, according to AFTE, "subjective in nature, based on one's training and experience."[303]

These ammunition identification techniques, like in other disciplines, are being increasingly automated. The current imaging system for identifying

299. *Theory of Identification, AFTE*, *supra* note 294, at 86.

300. Joseph L. Peterson et al., *Crime Laboratory Proficiency Testing Research Program*, Nat'l Inst. L. Enforcement & Crim. Just. 207 (1978), https://perma.cc/BDJ8-JRNP.

301. Gerald Burrard, The Identification of Firearms and Forensic Ballistics 107 (1962). However, bullet and cartridge case identifications differ in several respects. Because the bullet is traveling through the barrel at the time it is imprinted with the bore imperfections, these marks are "sliding" imprints, called striated marks. In contrast, the cartridge case receives "static" imprints, called impressed marks. *Id.* at 145.

302. Ammunition itself also has "class" characteristics, including markings on the head of the cartridge case, known as "head stamps," as well as bullet characteristics, such as caliber, type (such as hollow-point), weight, and jacketing, which may link cartridge cases or expelled projectiles at a scene to others collected elsewhere (such as unexpended cartridges within a seized firearm or at a search location), by caliber, brand, or type. *See, e.g.*, Defense Intelligence Agency, Small-Caliber Ammunition Identification Guide, Vol. 2 (1985), https://perma.cc/U4Q2-NXE9.

303. Eliot Springer, *Toolmark Examinations—A Review of Its Development in the Literature*, 40 J. Forensic Scis. 964, 966–67 (1995), https://doi.org/10.1520/JFS13864J.

bullets is the Integrated Ballistics Information System (IBIS).[304] Automated systems screen bullets and cartridge cases in the database for possible "matches."[305] IBIS identifies a number of candidate bullets, and the examiner makes a final comparison with a microscope.[306] The examiner has discretion to reject all the candidates.

The Toolmark Comparison Method and Its Claims

Toolmark comparisons rest on essentially the same theory as firearms identifications.[307] Toolmark comparisons rest on the assumption that tools, and the marks they leave on surfaces like wood or metal, have both class and individual characteristics.[308] For toolmarks, "class" characteristics might include the size of the mark/tool or the type of tool. For example, it may be possible to identify a mark (impression) left on wood as having been produced by a hammer, a knife, or a screwdriver. "Individual" characteristics are thought to include, for example, accidental imperfections produced by the machining process and subsequent use. When the tool is used, these characteristics are sometimes imparted onto the surface of another object struck by the tool.

As with firearms identification testimony, toolmark identification testimony is based on the largely subjective judgment of the examiner, who determines whether a sufficient number of marks of sufficient similarity are present to permit an identification.[309] There are no easily repeatable or reproducible steps involving objective criteria governing the determination of whether there is a "match."

304. *See* Jan De Kinder et al., *Reference Ballistic Imaging Database Performance*, 140 Forensic Sci. Int'l 207 (2004), https://doi.org/10.1016/j.forsciint.2003.12.002; Ruprecht Nennstiel & Joachim Rahm, *An Experience Report Regarding the Performance of the IBIS™ Correlator*, 51 J. Forensic Scis. 24 (2006), https://doi.org/10.1111/j.1556-4029.2005.00003.x.

305. Richard E. Tontarski & Robert M. Thompson, *Automated Firearms Evidence Comparison: A Forensic Tool for Firearms Identification—An Update*, 43 J. Forensic Scis. 641, 641 (1998), https://doi.org/10.1520/JFS16196J.

306. *Id.*

307. *See* Springer, *supra* note 303, at 964:

> The identification is based . . . on a series of scratches, depressions, and other marks which the tool leaves on the object it comes into contact with. The combination of these various marks ha[s] been termed toolmarks and the claim is that every instrument can impart a mark individual to itself.

308. While there are no studies assessing the phenomena of subclass characteristics for tools, in principle there is no reason to believe nongun tools would not also have subclass characteristics to the same extent as firearms.

309. *See* Springer, *supra* note 303, at 966–67 ("According to the Association of Firearms and Toolmarks Examiners' Criteria for Identification Committee, interpretation of toolmark individualization and identification is still considered to be subjective in nature, based on one's training and experience.").

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of firearms analysis,[310] given that these will frame the litigation judges will face over admission of firearm comparison testimony.

Critics of firearm identification raise three primary challenges. First, they argue that it is premised on an empirically untested assumption: that every firearm leaves unique, individualized markings on spent ammunition, a direct trail back to the firearm in question. Without variability data to back up this assumption, critics argue that claims of source attribution are not yet sustainable. Second, they argue that the method of comparison itself is overly subjective, based largely on examiner judgment and experience and thus prone to cognitive bias. Third, critics argue that the method is insufficiently validated through well-designed studies as an accurate means of establishing whether ammunition was fired from a particular firearm. In particular, they argue that existing studies suffer serious design flaws that underestimate error rate. Linking ammunition to a larger class of firearms is less controversial, though also sometimes challenged, as discussed below.

The Crime Laboratory Proficiency Testing Program, begun in 1978, was the first published test of the accuracy of firearms analysis.[311] In one test, 5.3% of the participating laboratories misidentified firearms evidence, and in another test 13.6% erred. These tests involved bullet and cartridge case comparisons. The Project Advisory Committee considered these errors "particularly grave in nature" and concluded that they probably resulted from carelessness, inexperience, or inadequate supervision.[312] A third test required the examination of two bullets and two cartridge cases to identify the "most probable weapon" from which each was fired. The error rate was 28.2%. In later tests,

> [e]xaminers generally did very well in making the comparisons. For all fifteen tests combined, examiners made a total of 2106 [bullet and cartridge case] comparisons and provided responses which agreed with the manufacturer responses 88% of the time, disagreed in only 1.4% of responses, and reported inconclusive results in 10% of cases.[313]

310. This section focuses on firearms identification, although it notes where studies relate to toolmarks as well. The broader issues (i.e., subjectivity, lack of variability data, lack of well-designed validation studies) are likely to arise in toolmark litigation as well.

311. Peterson et al., *supra* note 300.

312. *Id.* at 207–08.

313. Joseph L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, Part II: Resolving Questions of Common Origin*, 40 J. Forensic Scis. 1009, 1018 (1995), https://doi.org/10.1520/JFS13871J. The authors also stated:

> The performance of laboratories in the firearms tests was comparable to that of the earlier LEAA study, although the rate of successful identifications actually was slightly over—88% vs. 91%.

In this same report, toolmark identification had higher error rates than firearm identification.[314]

Citing this and other studies, the National Academy of Sciences in a 2008 Report on Ballistic Imaging concluded that "[t]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated."[315] Specifically, it found that "[m]ost of th[e existing firearms] studies are limited in scale and have been conducted by firearms examiners (and examiners in training) in state and local law enforcement laboratories as adjuncts to their regular casework."[316] As psychologists have later noted, the insular nature of tests makes the error rate difficult to ascertain from such tests alone.[317] In addition, noting that firearms examiners often testify to a definite identification, the 2008 report cautioned, "examiners tend to cast their assessments in bold absolutes, commonly asserting that a 'match' can be made 'to the exclusion of all other firearms in the world.' Such comments cloak an inherently subjective assessment of a 'match' with an extreme probability statement that has no firm grounding and unrealistically implies an error rate of zero."[318]

The 2009 NRC Report likewise reviewed the existing validation studies, concluding that the technique can narrow down to a class of guns or tools, but was not yet sufficiently validated for determining that a particular gun or tool is the source of a mark or pattern:

> Because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result. Sufficient studies have not been done to understand the reliability and repeatability of the methods. The committee agrees that class characteristics are helpful in narrowing the pool of tools that may have left a distinctive mark. Individual patterns from manufacture or from wear might, in some cases, be distinctive enough to suggest one

Laboratories cut the rate of errant identifications by half (3% to 1.4%) but the rate of inconclusive responses doubled, from 5% to 10%.

Id. at 1019.

314. *Id.* at 1025 ("Overall, laboratories performed not as well on the toolmark tests as they did on the firearms tests.").

315. National Research Council, Ballistic Imaging 81 (2008), https://doi.org/10.17226/12162.

316. *Id.* at 70. *See also* William A. Tobin, H. David Sheets & Clifford Spiegelman, *Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices*, 4 Statistics & Pub. Pol'y 1, 1 (2017), https://doi.org/10.1080/2330443X.2016.1270175 (stating that the majority of firearms testing that has been done has been developed by firearms/toolmark examiners, an "insular communit[y] of nonscientist practitioners . . . who did not incorporate effective statistical methods").

317. *See* Itiel E. Dror & Nicholas Scurich, *(Mis)use of Scientific Measurements in Forensic Science*, 2 Forensic Sci. Int'l: Synergy 333, 333 (2020), https://doi.org/10.1016/j.fsisyn.2020.08.006.

318. National Research Council, *supra* note 315, at 82.

> particular source, but additional studies should be performed to make the process of individualization more precise and repeatable.[319]

The report also identified a "fundamental problem with toolmark and firearms analysis" as being "the lack of a precisely defined process" or "specific protocol," which the report said led to a lack of "well-characterized confidence limits" in the examiners' conclusions.[320]

The 2016 PCAST Report has offered the most extensive analysis to date about *why* existing studies might underestimate error rate. The report explained that all but two existing firearms studies are either within-set studies or closed-set studies. In within-set studies, examiners analyze bullets to determine which came from the same firearm. The comparisons are not independent of each other (e.g., once bullet A matches bullet B, then if bullet C matches bullet A it must also match bullet B), which the report explained might artificially reduce the false-positive rate. In closed-set studies, examiners are asked to link each projectile to a gun in the set; all the right answers are included (all source guns are present). The problem with a closed-set study, according to the PCAST report, is that it is easier than real casework, for the same reason a multiple-choice question is easier (the right answer is definitely one of the choices given).[321]

PCAST identified two existing firearms studies—the "Miami Dade study" and the "Ames study" (by the Ames Laboratory, a Department of Energy laboratory affiliated with Iowa State University)—that were not within-set or closed-set. The Miami Dade study was conducted by the Miami Dade police crime laboratory and involved a "partial open-set" where at least some source guns were not among those present.[322] Among the conclusive determinations made, four were false positives, resulting in a false positive rate of 1 in 48, with an upper bound (of the likely range of the true false positive rate) of 1 in 19. If one includes inconclusive determinations in the denominator (which PCAST argued was inappropriate), the false positive rate is lower.[323] The Ames study was modeled after the FBI's latent print study, and involved true open sets, where the examinations were fully independent of each other and where there was no indication whether the source gun was present. In the Ames study, there were 22 false positives and 735 "inconclusives" out of 2,158 total "different source" examinations where the ground truth (the correct call, that is) was "exclusion."

319. 2009 NRC Report, *supra* note 37, at 154.

320. *Id.* at 155.

321. 2016 PCAST Report, *supra* note 20, at 107–09. The report notes that even where examiners are not explicitly told the study is a closed set, examiners can easily intuit this from the study design and answer sheet. *Id.* at 108.

322. *Id.* at 109. In the Miami Dade study, 165 examiners "were asked to assign a collection of 15 questioned samples, fired from 10 pistols to a collection of known standards; two of the 15 questioned samples came from a gun for which known standards were not provided. For these two samples, there were 188 eliminations, 138 inconclusives and 4 false positives." *Id.*

323. *Id.*

PCAST calculated the false positive rate from this information as being 22 out of 1,443 (the number of conclusive examinations), leading to an estimated error rate as high as 2.2% or 1 in 46.[324]

Ultimately, PCAST concluded that only the Ames study was sufficiently well designed to support a finding of foundational validity of the method, in that they involved designers who were independent of law enforcement (even though the participants themselves were self-selected)[325] and involved at least a partial open set. PCAST concluded that because of design flaws, other studies "seriously underestimate the false positive rate." With only one well-designed study to back it, firearms analysis in PCAST's view "still falls short of the scientific criteria for foundational validity."[326] PCAST urged that, if judges admitted firearms examination results in court, the testimony be accompanied by estimates of the false positive rate, based on the number of false positives in the examiners' total number of conclusive examinations.[327] The National District Attorneys Association disagreed with this conclusion in its response to PCAST, arguing that "the PCAST members are neither forensic firearm scientists performing casework nor did they participate as examiners in these validation studies" and that PCAST's criteria for a well-designed study was "arbitrarily defined."[328]

As judges might glean from the discussion above, there is a continuing debate about how to deal with "inconclusive" determinations in calculating a false positive rate. There are two separate issues. The first issue is whether to calculate a false positive rate based on the total number of examinations (including those where the examiner's opinion was "inconclusive") or solely based on conclusive examinations. The 2016 PCAST Report urged the latter approach, offering a hypothetical: "[C]onsider an extreme case in which a method had been tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/1,000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations)."[329]

The second issue is whether an "inconclusive" determination should be treated as an error when the ground truth is an "exclusion"—for example, where a bullet was definitively not fired from Gun A but the examiner fails to call this an exclusion. In one sense, this call is not the same as a false identification because the examiner has not stated that the bullet was fired from the gun. Moreover, in real casework, "inconclusive" might be the most accurate inference from the

324. *Id.* at 110.

325. Alan H. Dorfman & Richard Valliant, *Inconclusives, Errors, and Error Rates in Forensic Firearms Analysis: Three Statistical* Perspectives, 5 Forensic Sci. Int'l: Synergy 1, 5 (2022), https://doi.org/10.1016/j.fsisyn.2022.100273 (noting self-selecting volunteers for Ames study).

326. 2016 PCAST Report, *supra* note 20, at 111–12.

327. *Id.*

328. Nat'l Dist. Att'ys Ass'n, *supra* note 26, at 6.

329. 2016 PCAST Report, *supra* note 20, at 51–52.

evidence, given the conditions. In another sense, this call is erroneous in that it fails to exclude what is in fact an exclusion. This failure is especially concerning in a closed-set study where inconclusives are inherently a wrong response (because the source gun is always present), but is potentially concerning in nearly all validation studies, given that existing studies (unlike real casework) are designed to offer examiners sufficient information to correctly call exclusions when they are indeed exclusions.[330] As of this writing, "[r]esearchers have yet to agree on whether and how inconclusives should be used when assessing examiner performance."[331] The Food and Drug Administration, for its part, directs researchers facing this issue to calculate two error rates: one treating all inconclusives as a "positive" (here, an identification) and one treating inconclusives as a "negative" (here, an exclusion).[332] Judges should be aware of the debate and why "[i]t is clear that a well-founded characterization of inconclusives is critical for assessing the size of error rates estimated from the forensic firearms studies."[333]

Since the 2016 PCAST Report, several other "open-set" validation studies of firearms comparison have been published, some by the AFTE in its own journal[334] and others in peer-reviewed journals like the *Journal of Forensic Sciences*.[335] Some of these new studies have also been critiqued on grounds similar to

330. *See generally* Itiel E. Dror & Glenn Langenburg, *"Cannot Decide": The Fine Line Between Appropriate Inconclusive Determinations Versus Unjustifiably Deciding Not to Decide*, 64 J. Forensic Scis. 10, 11 (2018), https://doi.org/10.1111/1556-4029.13854 (arguing that inconclusives should be treated as error where ground truth is exclusion); Maneka Sinha & Richard Gutierrez, *Signal Detection Theory Fails to Account for Real-World Consequences of Inconclusive Decisions*, 21 L., Probability & Risk 131 (2023), https://doi.org/10.1093/lpr/mgad001 (same); Dror & Scurich, *supra* note 317, at 334 (same); Hal Arkes & Jonathan Jay Koehler, *Inconclusives and Error Rates in Forensic Science: A Signal Detection Theory Approach,* 20 L., Probability & Risk 153 (2021), https://doi.org/10.1093/lpr/mgac005 (arguing that whether inconclusives are error depends on the context, given that "inconclusive" is not a state of nature); Dorfman & Valliant, *supra* note 325, at 6–7 (noting that inconclusives should at least sometimes be treated as errors); *cf.* Andrew Smith & Gary Wells, *Telling Us Less than What They Know: Expert Inconclusive Reports Conceal Exculpatory Evidence in Forensic Cartridge-Case Comparisons*, 13 J. Applied Res. Memory & Cognition 147 (2024), https://doi.org/10.1037/mac0000138.

331. Arkes & Koehler, *supra* note 330.

332. *See* U.S. Food & Drug Admin., Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests 18 (2007), https://www.fda.gov/media/71147/download.

333. Dorfman & Valliant, *supra* note 325, at 2.

334. *See, e.g.*, Brandon A. Best & Elizabeth A. Gardner, *An Assessment of the Foundational Validity of Firearms Identification Using Ten Consecutively Button-Rifled Barrels*, 54 AFTE J. 28 (2022); Mark A. Keisler et al., *Isolated Pairs Research Study*, 50 AFTE J. 56 (2018).

335. *See, e.g.*, Eric F. Law & Keith B. Morris, *Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions*, 66 J. Forensic Scis. 1704 (2021), https://doi.org/10.1111/1556-4029.14758; Keith L. Monson, Eric D. Smith & Eugene M. Peters, *Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. Forensic Scis. 86 (2023), https://doi.org/10.1111/1556-4029.15152; Jaimie A. Smith, *Beretta Barrel Fired Bullet Validation Study*, 66 J. Forensic Scis. 547 (2021), https://doi.org/10.1111/1556-4029.14604. *See also* Max Guyll et al., *Validity of Forensic*

PCAST's critiques of the Ames study,[336] or because of their treatment of inconclusives[337] or other issues related to "missingness" of data.[338] Judges should be aware both that new studies are being published and of the recurring critiques of study design and error rate calculation that might be at the heart of the debate over such studies' probative value in determining reliability.

Reliability as applied. As with fingerprints and handwriting analysis, evidence of the reliability as applied of firearms and toolmark comparison will typically be a mix of proficiency testing (which is also related to expert qualification), internal validation of any software or equipment used, documentation of whether the examiner followed standards and protocols, documentation of steps taken to avoid contextual and other bias, and evidence about the condition of the particular sample being examined. As with other methods, a judge might find based on the evidence presented that firearms and toolmark analysis is reliable as applied to offering one type of conclusion or answering one type of question, but not another.

In terms of proficiency testing, one study cited by proponents of firearms comparison testimony suggests a low 1.4% error rate based on a 1978–1991 study of results of proficiency tests given by the main private forensic proficiency testing firm, Collaborative Testing Services (CTS).[339] In a 2005 CTS cartridge case examination, none of the 255 test-takers nationwide answered incorrectly.[340] As one district court judge has noted, "[o]ne could read these results to mean that

Cartridge-Case Comparisons, 120 Proc. Nat'l Acad. Scis. e2210428120 (2023), https://doi.org/10.1073/pnas.2210428120.

336. After the 2016 PCAST Report, the Ames Laboratory and FBI collaborated on a second study, known as Ames II, originally available online as Stanley J. Bajic et al., Ames Laboratory, U.S. Dep't of Energy, Technical Report No. ISTR-5220, *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearms Comparison* (Oct. 7, 2020). The published version of the study in the *Journal of Forensic Sciences*, however, includes only FBI authors. *See* Keith L. Monson et al., *Repeatability and Reproducibility of Comparison Decisions by Firearms Examiners*, 68 J. Forensic Scis. 1721 (2023), https://doi.org/10.1111/1556-4029.15318. The FBI authors now state, without further explanation: "We acknowledge the contracted research contributions of Ames Laboratory personnel, who have declined authorship and individual acknowledgment." *Id.* at 1738.

337. *See, e.g.*, Dror & Scurich, *supra* note 317, at 334 (discussing Keisler et al., *supra* note 334); Michael Rosenblum et al., *Commentary on Guyll et al. (2023): Misuse of Statistical Method Results in Highly Biased Interpretation of Forensic Evidence* 1 (Sept. 11, 2023), https://perma.cc/VB9J-DV2C (arguing that Guyll et al., *supra* note 335, "make a serious statistical error that leads to highly inflated claims about the probability that a cartridge case from a crime scene was fired from a reference gun").

338. *See, e.g.*, Khan & Carriquiry, *supra* note 102, at 4–5 (noting "missingness problems" affecting forensic error rate calculation, including self-selected participants, high rates of attrition, nonresponse, how inconclusive responses are used, low reproducibility (inter-examiner consistency) and repeatability (intra-examiner consistency)).

339. *See, e.g.*, United States v. Monteiro, 407 F. Supp. 2d 351, 367 (D. Mass. 2006) (noting that government was citing and relying on Richard Grzybowski et al., *Firearm/Toolmark Identification: Passing the Reliability Test Under Federal and State Evidentiary Standards*, 35 AFTE J. 209, 213 (2003))

340. *Monteiro*, 407 F. Supp. 2d at 367.

the technique is foolproof, but the results might instead indicate that the test was somewhat elementary."[341] Indeed, the CTS president has acknowledged that their customers prefer that tests are designed to be "easy."[342] Some laboratories, like the Houston Forensic Science Center, have moved to "blind" proficiency testing where examiners do not know they are being tested.[343] This solution has also been suggested as a means of accounting for artificially inflated accuracy rates where inconclusives are treated as non-errors.[344]

Case Law Development

Before 2005, courts had routinely allowed testimony of experts not only about class characteristics—or whether a bullet could have been fired from a particular firearm, or a mark could have been made by a particular tool[345]—but also more categorical statements,[346] other than where experts were attempting particularly novel comparisons.[347] In 2005, a federal district court judge ruled for the first time that an expert could not testify that recovered cases came from a

341. *Id.*

342. 2016 PCAST Report, *supra* note 20, at 57 n.133 (quoting CTS president).

343. Garrett & Mitchell, *supra* note 19, at 923.

344. *See, e.g.*, Arkes & Koehler, *supra* note 330 (suggesting blind proficiency testing as a possible solution to the inconclusive problem); Dorfman & Valliant, *supra* note 325, at 6–7 (same).

345. *E.g.*, People v. Horning, 102 P.3d 228, 236 (Cal. 2004) (expert "opined that both bullets and the casing could have been fired from the same gun . . . ; because of their condition he could not say for sure"); Luttrell v. Commonwealth, 952 S.W.2d 216, 218 (Ky. 1997) (expert "testified only that the bullets which killed the victim could have been fired from Luttrell's gun"); United States v. Murphy, 996 F.2d 94, 99 (5th Cir. 1993) (allowing testimony of FBI expert "that the tools such as the screwdriver associated with Murphy 'could' have made the marks on the ignitions but that he could not positively attribute the marks to the tools identified with Murphy").

346. *See, e.g.*, United States v. Bowers, 534 F.2d 186, 193 (9th Cir. 1976) (concluding that toolmark identification "rests upon a scientific basis and is a reliable and generally accepted procedure"); United States v. Hicks, 389 F.3d 514, 526 (5th Cir. 2004) (ruling that "the matching of spent shell casings to the weapon that fired them has been a recognized method of ballistics testing in this circuit for decades"); United States v. Foster, 300 F. Supp. 2d 375, 377 n.1 (D. Md. 2004) ("Ballistics evidence has been accepted in criminal cases for many years. . . . In the years since *Daubert,* numerous cases have confirmed the reliability of ballistics identification."); United States v. Santiago, 199 F. Supp. 2d 101, 111 (S.D.N.Y. 2002) ("The Court has not found a single case in this Circuit that would suggest that the entire field of ballistics identification is unreliable."); Commonwealth v. Whitacre, 878 A.2d 96, 101 (Pa. Super. Ct. 2005) ("no abuse of discretion in the trial court's decision to permit admission of the evidence regarding comparison of the two shell casings with the shotgun owned by Appellant").

347. *See, e.g.*, Ramirez v. State, 810 So. 2d 836, 849–51 (Fla. 2001) (rejecting testimony matching knife to a cartilage wound); Sexton v. State, 93 S.W.3d 96, 101 (Tex. Crim. App. 2002) (rejecting testimony matching collected fired cartridge cases to cases from *unfired* bullets found in appellant's apartment based solely on magazine marks).

specific weapon "to the exclusion of every other firearm in the world."[348] Other courts similarly began to disallow certain statements suggesting certainty or definitive source attribution.[349] Other courts continued to allow statements such as a firearm identification with "practical certainty."[350]

Several courts in the past decade have, for the first time, more significantly limited firearm comparison testimony to statements about class characteristics or statements that a gun "cannot be excluded" as being a source of a mark or

348. United States v. Green, 405 F. Supp. 2d 104 (D. Mass. 2005).

349. *See, e.g.*, Williams v. United States, 210 A.3d 734, 742 (D.C. 2019) (concluding that "the empirical foundation does not currently exist to permit these examiners to opine with certainty that a specific bullet can be matched to a specific gun," and that the trial court plainly erred in allowing the testimony); United States v. Diaz, No. CR 05-00167 WHA, 2007 WL 485967, at *1 (N.D. Cal. Feb. 12, 2007) (allowing examiner to testify that a match has been made to a "reasonable degree of certainty in the ballistics field" but not "to the exclusion of all other firearms in the world"); United States v. Glynn, 578 F. Supp. 2d 567, 568 (S.D.N.Y. 2008) (disallowing "reasonable degree of ballistic certainty" testimony but allowing examiner to say that gun was "more likely than not" the source); Gardner v. United States, 140 A.3d 1172, 1177 (D.C. 2016) (holding that "firearms and toolmark expert may not give an unqualified opinion, or testify with absolute or 100% certainty, that based on ballistics feature comparison matching a fatal shot was fired from one firearm, to the exclusion of all other firearms"); United States v. Willock, 696 F. Supp. 2d 536, 546, 549 (D. Md. 2010) (Holding, based on a comprehensive magistrate judge's report, that examiner "Sgt. Ensor shall not opine that it is a 'practical impossibility' for a firearm to have fired the cartridges other than the common 'unknown firearm' to which Sgt. Ensor attributes the cartridges." Thus, "Sgt. Ensor shall state his opinions and conclusions without any characterization as to the degree of certainty with which he holds them."); United States v. Taylor, 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009):

> [B]ecause of the limitations on the reliability of firearms identification evidence discussed above, Mr. Nichols will not be permitted to testify that his methodology allows him to reach this conclusion as a matter of scientific certainty. Mr. Nichols also will not be allowed to testify that he can conclude that there is a match to the exclusion, either practical or absolute, of all other guns. He may only testify that, in his opinion, the bullet came from the suspect rifle to within a reasonable degree of certainty in the firearms examination field.

See also United States v. Monteiro, 407 F. Supp. 2d 351, 374 (D. Mass. 2006) (allowing ballistics expert to testify to a "reasonable degree of certainty" but not a "statistical certainty").

350. *See, e.g.*, United States v. Williams, 506 F.3d 151, 161–62 (2d Cir. 2007) (allowing testimony of a ballistics "match" but cautioning that the opinion should not "be taken as saying that any proffered ballistic expert should be routinely admitted"); United States v. McCluskey, CR. No. 10-2734 JCH, 2013 WL 12335325, at *9–10 (D.N.M. Feb. 7, 2013) (allowing "practical certainty" but disallowing "reasonable degree of ballistic certainty" as a nonsensical nonscientific term); United States v. Natson, 469 F. Supp. 2d 1253, 1261 (M.D. Ga. 2007) (allowing statement of "100% degree of certainty"); Commonwealth v. Meeks, Nos. 2002-10961, 2003-10575, 2006 WL 2819423, at *50 (Mass. Super. Ct. Sept. 28, 2006) (allowing opinion of a "match"); United States v. Casey, 928 F. Supp. 2d 397, 399–400 (D.P.R. 2013) (finding that although "defendant challenges [the] conclusion that [the examiner] is 100% certain," the court "remains faithful to the long-standing tradition of allowing the unfettered testimony of qualified ballistics experts"); State v. Davidson, 509 S.W.3d 156, 205 (Tenn. 2016) ("it's like a fingerprint"); State v. Anderson, 624 S.E.2d 393, 397–98 (N.C. Ct. App. 2006) (no abuse of discretion in admitting bullet identification evidence).

pattern. In the often-cited case of *United States v. Tibbs,* a D.C. trial judge conducted a multiday evidentiary hearing, ultimately restricting the firearms testimony to a statement that the gun "cannot be excluded as the source of the cartridge case found on the scene," rather than identification, because "[a]ny statement by the expert involving more certainty regarding the relationship between a casing and a firearm would stray into territory not presently supported by reliable principles and methods."[351] The court recognized the "threshold design issues" of existing firearms/toolmark studies, which "limit their utility."[352] A leading scientific evidence treatise calls *Tibbs* the "a high water mark for genuinely engaging with the imperfect and limited state of the scientific research underlying firearms identification research,"[353] and several other courts since 2019 have similarly limited firearms examiner testimony.[354] Indeed,

351. *See* United States v. Tibbs, 2016-CF1-19431, 2019 WL 4359486, at *23 (D.C. Super. Ct. Sept. 5, 2019).

352. *Id.* at *14.

353. Modern Scientific Evidence, *supra* note 138, at § 34:5.

354. *See, e.g.*, United States v. Briscoe, No. 20-CR-1777 MV, 2023 WL 8096886 (D.N.M. Nov. 21, 2023) (citing *Tibbs* and allowing testimony but declining to allow evidence of "consistent with," "sufficient agreement," "match," but allowing class characteristics); Abruquah v. State, 296 A.3d 961, 996–98 (Md. 2023) (allowing only testimony that gun was "consistent or inconsistent" with markings); United States v. Cloud, 576 F. Supp. 3d 827, 845 (E.D. Wash. 2021):

> [I]f [examiner] intends to go beyond testimony that merely notes the recovered cartridge casings could not be excluded as having been fired from the recovered firearm, the Court will inform the jury that: (1) only three studies that meet the minimum design standard have attempted to measure the accuracy of firearm/toolmark comparison and (2) these studies found false positive rates that could be as high as 1 in 46 in one study, 1 in 200 in the second study, and 1 in 67 in the third study, though this study has yet to be published and subjected to peer review.

See also United States v. Shipp, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (ruling that ballistics expert would be permitted to testify only that toolmarks on recovered bullet fragment and shell casing were consistent with having been fired from the recovered firearm); United States v. Adams, 444 F. Supp. 3d 1248, 1267 (D. Or. 2020) (limiting testimony to shared class characteristics only, such as caliber, type of firing pin, "barrel with six lands/grooves and right twist"); People v. Ross, 129 N.Y.S.3d 629, 642 (Sup. Ct. 2020) (limiting testimony to class characteristics only, and that the gun "cannot be ruled out" as the source, but disallowing phrases like "consistent with" or "sufficient agreement"); Illinois v. Rickey Winfield, 15 CR 14066-01 (Cook Cty. Cir. Ct. Feb. 8, 2023) (excluding ballistics testimony, concluding "[t]here are no objective forensic based reasons that firearms identification evidence belongs in any category of forensic science" and that "the junk evidence" "fails" the *Frye* test for admissibility (Frye v. United States, 293 F. 1013 (D.C. Cir. 1923))); State v. Ghigliotty, No. 17-0200154-1, slip op. at 39 (N.J. Super. Ct. Law Div. Aug. 23, 2019):

> [T]his Court finds that the current state of firearms identification through the use of tool marks in the view of the scientific community does not support permitting [the expert] to frame his opinion that the identification of the evidence bullet in this case was "positive" or a "match" in terms of absolutely certainty.

Cf. United States v. Medley, No. PWG 17-242 (D. Md. Apr. 24, 2018) (pre-*Tibbs* opinion ruling that expert may not testify to a "match" opinion).

at least one court has cited *Tibbs* in rejecting a *defense* ballistics expert.[355] At the same time, however, other courts continued to uphold admission.[356]

As explained at the end of the sections titled "Fingerprint Evidence" and "Handwriting Evidence," the 2023 amendment to Rule 702 may require courts to engage in more robust gatekeeping to evaluate whether the opponent has proven by a preponderance of the evidence both the foundational reliability and the reliability as applied of firearm and toolmark comparison expert testimony.

Bitemark Evidence

Introduction

Forensic odontology is the field engaged in "examining, interpreting, and presenting dental and oral evidence for investigative and legal purposes."[357] Forensic bitemark examiners, a subset of forensic odontologists, purport to be able to identify a mark as a human bitemark and then attribute it to a particular person's teeth.[358] Other forensic odontologists focus only on human identification—that is, identifying people by their teeth.

Courts previously admitted bitemark analysis without much scrutiny into its reliability. In more recent years, scientists have overwhelmingly concluded that bitemark evidence is not reliable.[359] As discussed below, research has

355. *See* United States v. Harris, 491 F. Supp. 3d 414, 424–25 (E.D. Wis. 2020) (rejecting on *Daubert* grounds proffered defense expert testimony that mark on car was not from a bullet, citing *Tibbs*).

356. *See, e.g.*, United States v. Hunt, 63 F.4th 1229, 1249, n.18 (10th Cir. 2023) (allowing testimony based on CMS (consecutive matching striae) and noting that *Tibbs* did not deal with that method); United States v. Perry, 35 F.4th 293, 329–31 & n.23 (5th Cir. 2022) (holding that trial court's admission of ballistics testimony under *Daubert* was not an abuse of discretion, notwithstanding the defendant's arguments as to lack of peer review and standardization); United States v. Graham, No. 4:23-CR-00006, 2024 WL 688256, at *16 (W.D. Va. Feb. 20, 2024) (denying *Daubert* challenge to ballistics expert, but requiring expert to testify consistent with FBI's new ULTR); United States v. Blackman, No. 18-CR-00728, 2023 WL 3440384, at *5 n.5, *9 (N.D. Ill. May 12, 2023) (allowing source attribution testimony so long as expert does not imply absolute certainty or "to the exclusion of all other" guns) (citing United States v. Green, 405 F. Supp. 2d 104, 124 (D. Mass. 2005)); United States v. Harris, 502 F. Supp. 3d 28, 33 (D.D.C. 2020) (denying *Daubert* challenge to ballistics expert, but requiring expert to testify consistent with FBI's new ULTR).

357. NIST, *Forensic Odontology Subcommittee*, https://perma.cc/5L4U-HTG6 (last updated Nov. 3, 2023).

358. *See* E.H. Dinkel, Jr., *The Use of Bite Mark Evidence as an Investigative Aid*, 19 J. Forensic Scis. 535 (1974), https://doi.org/10.1520/JFS10208J.

359. NIST Bitemark Report, *supra* note 110. As the Texas Forensic Science Commission wrote, "if anyone should take responsibility for the current state of bite mark comparison, it is the very organization of practitioners that, due to its glacial pace, reticence to publish critical data, and willingness to allow overstatements of science to go unchecked for decades, is facing a barrage of well-founded criticism." Tex. Forensic Sci. Comm'n, Forensic Bitemark Comparison Complaint

documented that (1) skin is an unstable medium upon which to expect consistent imprint results; (2) dental molds are frequently similar in nature and challenging to distinguish; and (3) bitemark experts themselves have increasing difficulty determining first whether a blemish is a bitemark, then whether it is a human or animal bitemark, and finally, whether the bitemark "matches" to a dental mold. The uniqueness of dentition has also never been proven. The OSAC Forensic Odontology Subcommittee itself makes clear on its website that it "does not develop standards on bitemark recognition, comparison, and identification."[360]

Still, as of this writing, some prosecutors and defense attorneys continue to offer bitemark evidence, and some courts continue to admit it.

The Method and Its Claims

Human identification. Forensic odontologists identify who a deceased person might be from human remains by comparing the decedent's teeth to dental records of a known person. The human adult set of teeth (dentition) offers many points of comparison for these purposes. A set of teeth consists of thirty-two teeth, each with five anatomic surfaces. Restorations, with varying shapes, sizes, and restorative materials, may offer numerous additional points of comparison. Moreover, the number of teeth, prostheses, decay, malposition, malrotation, peculiar shapes, root canal therapy, bone patterns, bite relationship, and oral pathology may also provide points of comparison.[361]

Although the assumption that every person's full dentition is unique remains empirically unproven,[362] "the identification of human remains by their dental characteristics is" nonetheless "well established."[363] The reliability of human identification through dental records stems from the finite number of candidates to identify and the availability of full dentition on both remains and records.[364]

Filed by National Innocence Project on behalf of Steven Mark Chaney—Final Report, 1, 17 (Apr. 12, 2016), https://www.txcourts.gov/media/1454500/finalbitemarkreport.pdf.

360. NIST, *supra* note 357.

361. The identification is made by comparing the decedent's teeth with antemortem dental records. *See generally* Javier Ata-Ali & Fadi Ata-Ali, *Forensic Dentistry in Human Identification: A Review of the Literature*, 6 J. Clin. & Exp. Dent. e162 (2014), https://doi.org/10.4317/jced.51387 (explaining points of comparison, process, and types of records relied on).

362. *See* 2009 NRC Report, *supra* note 37, at 175 ("The uniqueness of the human dentition has not been scientifically established."); NIST Bitemark Report, *supra* note 110, at i (finding a "lack of support" for this "key premise").

363. 2009 NRC Report, *supra* note 37, at 173. *See also* Satomi Mizuno et al., *Validity of Dental Findings for Identification by Postmortem Computed Tomography*, 341 Forensic Sci Int'l 111507 (2022), https://doi.org/10.1016/j.forsciint.2022.111507.

364. *See* Iain A. Pretty & David J. Sweet, *The Scientific Basis for Human Bitemark Analyses—A Critical Review,* 41 Sci. & Just. 85, 88 (2001), https://doi.org/10.1016/S1355-0306(01)71859-X

Bitemark identification. Bitemark identification, by contrast to human remains identification, relies on impressions and bruising of the skin and comparing those to a suspect's teeth. In contrast to human remains identification, the number of potential suspects may be large, the examiner may have only a small number of impressions to work with, the marks may change over time, and the medium on which the marks are observed (like skin) may be unstable and allow shrinkage and distortion.[365] Moreover, all five anatomic surfaces are not engaged in biting; only the edges of the front teeth come into play. In sum, bitemark identification depends not only on the uniqueness of each person's dentition but also on "whether there is a [sufficient] representation of that uniqueness in the mark found on the skin or other inanimate object."[366] This uniqueness has yet to be proven.[367]

Of the several methods of bitemark analysis that have been reported, all involve three steps: (1) registration of both the bitemark and the suspect's dentition, (2) comparison of the dentition and bitemark, and (3) evaluation of the points of similarity or dissimilarity. The comparison may be either direct or indirect. A model of the suspect's teeth is used in direct comparisons; the model is compared to life-size photographs of the bitemark. Transparent overlays made from the model are used in indirect comparisons.[368] The ultimate opinion of a bitemark examiner regarding individuation is a relatively subjective one.[369] For example, there is no accepted minimum number of points of identity required

("A distinction must be drawn from the ability of a forensic dentist to identify an individual from their dentition by using radiographs and dental records and the science of bitemark analysis.").

365. *See* 2009 NRC Report, *supra* note 37, at 174–76 (noting these limitations).

366. Raymond D. Rawson et al., *Statistical Evidence for the Individuality of the Human Dentition*, 29 J. Forensic Scis. 245, 252 (1984), https://doi.org/10.1520/JFS11656J.

367. Mary A. Bush et al., *Statistical Evidence for the Similarity of the Human Dentition*, 56 J. Forensic Scis. 118, 118 (2010), https://doi.org/10.1111/j.1556-4029.2010.01531.x (observing significant correlations and nonuniform distributions of tooth positions as well as "matches" between dentitions and concluding that "statements of dental uniqueness with respect to bitemark analysis in an open population are unsupportable and that use of the product rule is inappropriate"); Mary A. Bush et al., *Similarity and Match Rates of the Human Dentition in Three Dimensions: Relevance to Bitemark Analysis*, 125 Int'l J. Legal Med. 779, 779 (2011), https://doi.org/10.1007/s00414-010-0507-8 (explaining that three dimensional models reduce but do not limit random matches and "a zero match rate cannot be claimed for the population studied"); H. David Sheets et al., *Dental Shape Match Rates in Selected and Orthodontically Treated Populations in New York State: A Two-Dimensional Study*, 56 J. Forensic Sci. 621, 621 (2011), https://doi.org/10.1111/j.1556-4029.2011.01731.x (finding random dental shape "matches" and concluding that "statements of certainty concerning individualization in such populations should be approached with caution").

368. *See* David J. Sweet, *Human Bitemarks: Examination, Recovery, and Analysis, in* Manual of Forensic Odontology 162 (American Society of Forensic Odontology, 3d ed. 1997) ("The analytical protocol for bitemark comparison is made up of two broad categories. Firstly, the measurement of specific traits and features called a metric analysis, and secondly, the physical matching or comparison of the configuration and pattern of the injury called a pattern association.").

369. *See* Roland F. Kouble & Geoffrey T. Craig, *A Comparison Between Direct and Indirect Methods Available for Human Bite Mark Analysis*, 49 J. Forensic Scis. 111, 111 (2004), https://doi.org/10.1520/JFS2001252 ("It is important to remember that computer-generated overlays still retain an

for a positive identification,[370] and experts have testified to a wide range of points of similarity, from a low of eight points to a high of fifty-two.[371]

In an attempt to develop a more objective method, in 1984 the American Board of Forensic Odontology (ABFO) promulgated guidelines for bitemark analysis, including a uniform scoring system.[372] According to the drafting committee, "[t]he scoring system . . . has demonstrated a method of evaluation that produced a high degree of reliability among observers."[373] Moreover, the committee characterized "[t]he scoring guide . . . [as] the beginning of a truly scientific approach to bite mark analysis."[374] In a subsequent letter, however, the drafting committee wrote:

> While the Board's published guidelines suggest use of the scoring system, the authors' present recommendation is that all odontologists await the results of further research before relying on precise point counts in evidentiary proceedings. . . . [T]he authors believe that further research is needed regarding the quantification of bite mark evidence before precise point counts can be relied upon in court proceedings.[375]

Ultimately, the ABFO's effort to develop a standardized bitemark methodology "failed . . . due to inter examiner discord and unreliable quantitative interpretation."[376] The ABFO has conceded that bitemark evidence cannot be used for individualization in "open population" cases, where there is an unknown number of potential suspects.[377]

element of subjectivity, as the selection of the biting edge profiles is reliant on the operator placing the 'magic wand' onto the areas to be highlighted within the digitized image.").

370. *See* Stubbs v. State, 845 So. 2d 656, 669 (Miss. 2003) ("There is little consensus in the scientific community on the number of points which must match before any positive identification can be announced."). Leigh Stubbs was exonerated in 2012. *See* Valena Beety, Manifesting Justice: Wrongly Convicted Women Reclaim Their Rights (2022).

371. *E.g.*, State v. Garrison, 585 P.2d 563, 566 (Ariz. 1978) (10 points); People v. Slone, 143 Cal. Rptr. 61, 67 (Ct. App. 1978) (10 points); People v. Milone, 356 N.E.2d 1350, 1356 (Ill. App. Ct. 1976) (29 points); State v. Sager, 600 S.W.2d 541, 564 (Mo. Ct. App. 1980) (52 points); State v. Green, 290 S.E.2d 625, 630 (N.C. 1982) (14 points); State v. Temple, 273 S.E.2d 273, 279 (N.C. 1981) (8 points); Kennedy v. State, 640 P.2d 971, 976 (Okla. Crim. App. 1982) (40 points); State v. Jones, 259 S.E.2d 120, 125 (S.C. 1979) (37 points).

372. Am. Board of Forensic Odontology (ABFO), *Guidelines for Bite Mark Analysis*, 112 J. Am. Dental Ass'n 383 (1986).

373. Raymond D. Rawson et al., *Reliability of the Scoring System of the American Board of Forensic Odontology for Human Bite Marks*, 31 J. Forensic Scis. 1235, 1256 (1986), https://doi.org/10.1520/JFS11903J.

374. *Id.* at 1259.

375. Gerald L. Vale et al., *Letters to the Editor: Discussion of "Reliability of the Scoring System of the American Board of Forensic Odontology for Human Bite Marks,"* 33 J. Forensic Scis. 20 (1988).

376. C. Michael Bowers, *Problem-Based Analysis of Bitemark Misidentifications: The Role of DNA*, 159S Forensic Sci. Int'l S104, S106 (2006), https://doi.org/10.1016/j.forsciint.2006.02.032.

377. Am. Bd. of Forensic Odontology, Inc., Diplomates Reference Manual 102 (2015) (explaining that "[t]he ABFO does not support a conclusion of 'The Biter' in an open population case(s)").

Scientific Assessments, Critiques, and Debates

Bitemark analysis is rarely used in the United States because of serious concerns about its reliability, in part inspired by high-profile DNA exonerations involving people convicted based on bitemark testimony. For example, in *State v. Krone*,[378] two experienced experts concluded that the defendant had made the bitemark found on a murder victim. The defendant, however, was later exonerated through DNA testing.[379] In *Otero v. Warnick*,[380] a forensic dentist testified that the

> plaintiff was the only person in the world who could have inflicted the bite marks on [the murder victim's] body. On January 30, 1995, the Detroit Police Crime Laboratory released a supplemental report that concluded that plaintiff was excluded as a possible source of DNA obtained from vaginal and rectal swabs taken from [the victim's] body.[381]

In *Burke v. Town of Walpole*,[382] the expert concluded that "Burke's teeth matched the bite mark on the victim's left breast to a 'reasonable degree of scientific certainty.' That same morning . . . DNA analysis showed that Burke was excluded as the source of male DNA found in the bite mark on the victim's left breast."[383] These are but a few of the exonerations involving false bitemark evidence.

The 2009 NRC Report concluded that neither "[t]he uniqueness of human dentition," "[t]he ability of the dentition, if unique, to transfer a unique pattern to human skin and the ability of the skin to maintain that uniqueness" nor any "standard for the type, quality, and number of individual characteristics required to indicate that a bite mark has reached a threshold of evidentiary value" has been scientifically established.[384] Moreover, as the report noted, "[t]here is no science on the reproducibility of the different methods of analysis that lead to conclusions about the probability of a match. This includes reproducibility between experts and with the same expert over time. Even when using the

378. 897 P.2d 621, 622, 623 (Ariz. 1995) ("The bite marks were crucial to the State's case because there was very little other evidence to suggest Krone's guilt."; "Another State dental expert, Dr. John Piakis, also said that Krone made the bite marks . . . Dr. Rawson himself said that Krone made the bite marks. . . .").

379. *See* Mark Hansen, *The Uncertain Science of Evidence*, A.B.A. J., July 28, 2005, at 49, https://perma.cc/G674-XKPB (discussing *Krone*).

380. 614 N.W.2d 177 (Mich. Ct. App. 2000).

381. *Id.* at 178.

382. 405 F.3d 66 (1st Cir. 2005).

383. *Id.* at 73. *See also* Bowers, *supra* note 376, at S106–07 (citing several cases involving bitemarks and DNA exonerations); Mark Hansen, *Out of the Blue*, A.B.A. J., Feb. 1, 1996, at 50, 51, https://perma.cc/GV32-SAC4 (DNA analysis of skin taken from fingernail scrapings of the victim conclusively excluded suspect).

384. 2009 NRC Report, *supra* note 37, at 175–76. *See also id.* at 176 ("Although the majority of forensic odontologists are satisfied that bite marks can demonstrate sufficient detail for positive identification, no scientific studies support this assessment, and no large population studies have been conducted.").

[American Board of Forensic Odontology] guidelines, different experts provide widely differing results and a high percentage of false positive matches of bite marks using controlled comparison studies."[385] As as result, as early as 2009, there was already "continuing dispute over the value and scientific validity of comparing and identifying bite marks."[386] Indeed, because of the dispute, some odontologists well before the 2009 NRC Report had argued that "bitemark evidence should only be used to exclude a suspect."[387]

In 2015, the ABFO conducted an internal study titled *Construct Validity of Bitemark Assessments Using the ABFO Bitemark Decision Tree*.[388] This experiment presented 100 injuries to 39 ABFO board-certified forensic odontologists, to determine whether the injury at issue was a human bitemark and, if so, whether it had distinct identifiable arches and individual toothmarks.[389] Thirty-nine forensic odontologists completed the survey and "came to unanimous agreement on just 4 of the 100 case studies. Of the initial 100, there remained just 8 case studies in which at least 90 percent of the analysts were still in agreement."[390]

The Texas Forensic Science Commission, in a 2016 report, concluded that "there is no scientific basis for stating that a particular patterned injury can be associated to an individual's dentition" and "[a]ny testimony describing human dentition as 'like a fingerprint' or incorporating similar analogies lacks scientific support."[391] It ultimately urged that analyst testimony regarding the probability or weight of any association between a bitemark and an individual's dentition has "no place in our criminal justice system because they lack any credible supporting data."[392]

The 2016 PCAST Report likewise concluded that "available scientific evidence strongly suggests that examiners not only cannot identify the source of bitemark with reasonable accuracy, they cannot even consistently agree on whether an injury *is* a human bitemark."[393]

Most recently, in 2023, NIST released a report, *Bitemark Analysis: A NIST Scientific Foundation Review*, reviewing bitemark analysis. The report found

> forensic bitemark analysis lacks a sufficient scientific foundation because the three key premises of the field are not supported by the data. First, human

385. *Id.* at 174.

386. *Id.* at 173.

387. Iain A. Pretty, *A Web-Based Survey of Odontologist's [sic] Opinions Concerning Bitemark Analyses*, 48 J. Forensic Scis. 1117, 1120 (2003), https://doi.org/10.1520/JFS2003017.

388. Adam J. Freeman & Iain A. Pretty, Construct Validity of Bitemark Assessments Using the ABFO Decision Tree (2015), https://perma.cc/D7FZ-JAJX.

389. Radley Balko, *A Bite Mark Matching Advocacy Group Just Conducted a Study that Discredits Bite Mark Evidence*, Wash. Post, Apr. 8, 2015, https://perma.cc/S2DL-ZREX.

390. *Id.*

391. Tex. Forensic Sci. Comm'n, *supra* note 359, at 11–12.

392. *Id.* at 12.

393. 2016 PCAST Report, *supra* note 20, at 9 (emphasis in original).

anterior dental patterns have not been shown to be unique at the individual level. Second, those patterns are not accurately transferred to human skin consistently. Third, it has not been shown that defining characteristics of those patterns can be accurately analyzed to exclude or not exclude individuals as the source of a bitemark.[394]

Case Law Development

With respect to the use of dentition to identify human remains, courts have readily accepted the method as a means of establishing the identity of a homicide victim,[395] with some cases dating back to the nineteenth century.[396]

With respect to bitemarks, *People v. Marx*[397] emerged as the seminal case. The case "came to be read as a global warrant to admit bite mark identification evidence whenever a person displaying apparent credentials chose to testify to an identification."[398] The cases that closely followed and relied on *Marx* also went to great lengths to extoll the "superior trustworthiness of the scientific bite mark approach."[399] After *Marx*, bitemark evidence became widely accepted.[400] By 1992, it had been introduced or noted in 193 reported cases and accepted as

394. NIST Bitemark Report, *supra* note 110, at 24.

395. *E.g.*, Wooley v. People, 367 P.2d 903, 905 (Colo. 1961) (dentist compared his patient's record with dentition of a corpse); Martin v. State, 636 N.E.2d 1268, 1272 (Ind. Ct. App. 1994) (dentist qualified to compare X-rays of one of his patients with skeletal remains of murder victim and make a positive identification); Fields v. State, 322 P.2d 431, 446 (Okla. Crim. App. 1958) (murder case in which victim was burned beyond recognition).

396. *See* Commonwealth v. Webster, 59 Mass. 295, 299–300 (1850) (remains of the incinerated victim, including charred teeth and parts of a denture, were identified by the victim's dentist); Lindsay v. People, 63 N.Y. 143, 145–46 (1875).

397. 126 Cal. Rptr. 350 (Cal. Ct. App. 1975). The court in *Marx* avoided applying the *Frye* test, which requires acceptance of a novel technique by the scientific community as a prerequisite to admissibility. *See* Frye v. United States, 293 F. 1013 (D.C. Cir. 1923). According to the *Marx* court, the *Frye* test "finds its rational basis in the degree to which the trier of fact must accept, on faith, scientific hypotheses not capable of proof or disproof in court and not even generally accepted outside the courtroom." 126 Cal. Rptr. at 355–56.

398. D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock?*, 64 Alb. L. Rev. 99, 138 (2000).

399. People v. Slone, 143 Cal. Rptr. 61, 69 (Ct. App. 1978); *see also* State v. Sager, 600 S.W.2d 541, 569 (Mo. Ct. App. 1980) (characterizing bitemark evidence as "an exact science").

400. Two Australian cases, however, excluded bitemark evidence. *See* Lewis v The Queen (1987), 29 A Crim R 267 (odontological evidence was improperly relied on, in that this method has not been scientifically accepted); R v Carroll (1985), 19 A Crim R 410 ("[T]he evidence given by the three odontologist is such that it would be unsafe or dangerous to allow a verdict based upon it to stand.").

admissible in 35 states.[401] Some courts described bitemark comparison as an "exact science,"[402] and several cases took judicial notice of its validity.[403] The first state supreme court to take judicial notice of the "general acceptance" of bitemark evidence was West Virginia in *State v. Armstrong*, prompting other state supreme courts to rule similarly.[404] *State v. Armstrong*, however, relied on the findings in another bitemark identification case, that of Robert Lee Stinson in Wisconsin.[405] Stinson was ultimately exonerated by DNA evidence in 2009.[406] Hence, even the bitemark identification in the case undergirding what led to the national "general acceptance" of bitemark evidence was wrong.

Because law enforcement around the country has so significantly curtailed its reliance on bitemark evidence, there is little recent case law on the subject. Indeed, serious court evaluation of bitemarks has still been largely confined to civil postconviction habeas corpus cases and 42 U.S.C. § 1983 lawsuits for wrongful conviction and presentation of false evidence at trial.[407] Still, some courts have continued to admit the evidence.[408]

401. Steven Weigler, *Bite Mark Evidence: Forensic Odontology and the Law*, 2 Health Matrix: J.L.-Med. 303, 303 (1992).

402. *See* People v. Marsh, 441 N.W.2d 33, 35 (Mich. Ct. App. 1989) ("the science of bite mark analysis has been extensively reviewed in other jurisdictions"); *Sager*, 600 S.W.2d at 569 ("an exact science").

403. *See* State v. Richards, 804 P.2d 109, 112 (Ariz. Ct. App. 1990) ("[B]ite mark evidence is admissible without a preliminary determination of reliability. . . ."); People v. Middleton, 429 N.E.2d 100, 101 (N.Y. 1981) ("The reliability of bite mark evidence as a means of identification is sufficiently established in the scientific community to make such evidence admissible in a criminal case, without separately establishing scientific reliability in each case. . . ."); State v. Armstrong, 369 S.E.2d 870, 877 (W. Va. 1988) (judicially noticing the reliability of bitemark evidence).

404. *Armstrong*, 369 S.E.2d at 877.

405. State v. Stinson, 397 N.W.2d 136 (Wis. Ct. App. 1986).

406. *See* Jennifer D. Oliva & Valena E. Beety, *Discovering Forensic Fraud*, 112 Nw. U. L. Rev. 121 (2017).

407. *See, e.g.*, Keko v. Hingle, 318 F.3d 639, 644 (5th Cir. 2003) (denying absolute immunity to forensic odontologist in § 1983 civil lawsuit following wrongful conviction); Ege v. Yukins, 380 F. Supp. 2d 852, 871 (E.D. Mich. 2005), *aff'd in part, rev'd in part on other grounds*, 485 F.3d 364 (6th Cir. 2007) (ruling "there is no question that the [bitemark] evidence in this case was unreliable and not worthy of consideration by a jury"); *In re* Richards, 371 P.3d 195, 211 (Cal. 2016) (court granting civil writ of habeas corpus ruling bitemark expert's criminal trial testimony constituted material false evidence); Stinson v. Milwaukee, No. 09-C-1033, 2013 WL 5447916, at *12–13, *15–17 (E.D. Wis. Sept. 30, 2013), *aff'd in part, rev'd in part sub nom.* Stinson v. Gauger, 799 F.3d 833 (7th Cir. 2015) (denying absolute immunity to forensic odontologists in § 1983 civil lawsuit alleging fabrication and suppression of evidence).

408. *See* Aliza B. Kaplan & Janis C. Puracal, *It's Not a Match: Why the Law Can't Let Go of Junk Science*, 81 Alb. L. Rev. 895, 898 (2018) (notes courts that continue to admit bitemarks, relying on precedent); Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J.L. & Biosciences 538, 546–47 (2016), https://doi.org/10.1093/jlb/lsw045 (same).

Facial Recognition Software and Other Methods

Several other forensic feature comparison techniques are on the horizon, beyond the capacity of this reference guide to discuss. For example, "microbial forensics" is a biometric comparison technique that is increasingly used where DNA typing is not possible.[409] Additionally, digital forensics—from cell phone extraction to cell site location information to hash-value authentication of documents—is a significant area of *Daubert* litigation, although digital forensics is not typically referred to as a feature comparison technique. Digital evidence is discussed in the *Reference Guide on Computer Science*, in this manual.

One technique that has been largely supplanted by mitochondrial DNA analysis, and yet which still arises in federal habeas litigation, is microscopic hair comparison. In this technique, the expert compares hair samples under a microscope and testifies whether the hair samples share a common source. Microscopic hair analysis was described by the 2009 NRC Report as a field in which there is no uniform standard on the number of features that must be congruent between hair samples for an examiner to claim a "match." The report found that microscopic hair analysis without mtDNA testing was unreliable.[410] Likewise, the 2016 PCAST Report found insufficient evidence of foundational validity to support microscopic hair analysis.[411]

After the release of the 2009 NRC Report, the FBI launched a national reopening of closed cases involving hair analysis, and acknowledged misleading testimony by agents.[412] Of the first 200 cases internally reviewed by the FBI, over 90% contained false or faulty testimony.[413] Most notably, Santae Trible was imprisoned for twenty-eight years when FBI forensic analysts confused Trible's hair with a dog's hair.[414] In partnership with the DOJ, the National Association of Criminal Defense Attorneys, and the Innocence Project, the FBI agreed to

409. *See, e.g.*, Audrey Gouello et al., *Analysis of Microbial Communities: An Emerging Tool in Forensic Sciences*, 12 Diagnostics 1 (2021), https://doi.org/10.3390/diagnostics12010001 (discussing the techniques for identifying the anatomical origin of a microbial trace, and the forensic possibilities for different types of microbial evidence).

410. *See* 2009 NRC Report at 161, *supra* note 37.

411. *See* 2016 PCAST Report, *supra* note 20.

412. *See, e.g.*, Press Release, Innocence Project, *Innocence Project and NADCL Announce Historic Partnership with the FBI and Department of Justice on Microscopic Hair Analysis Cases* (July 18, 2013), https://perma.cc/35VU-ALVD.

413. Norman L. Reimer, *The Hair Microscopy Review Project: An Historic Breakthrough for Law Enforcement and a Daunting Challenge for the Defense Bar*, Champion, July 2013, at 16, https://perma.cc/679W-VS4T. *See also* Press Release, FBI, *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90% of Cases in Ongoing Review* (Apr. 20, 2015), https://perma.cc/Y3R4-3RV8.

414. *See* Simon A. Cole et al., *Microscopic Hair Comparison Analysis and Convicting the Innocent,* Nat'l Registry of Exonerations (Dec. 2023), https://perma.cc/WD7V-EDRD.

provide free DNA testing of any remaining evidence in the acknowledged cases. The DOJ agreed to waive any statute of limitation barriers. State initiatives followed in its wake.[415]

Finally, a few notes about facial recognition methods are in order, given their already wide investigative use and the fact that they raise issues similar to those raised by older, relatively subjective non-DNA feature comparison methods. Facial recognition software (or facial recognition technology, FRT) compares two images to determine whether the same person is present in each image.[416] A probe photo, such as a still from a surveillance video, can be uploaded to a police database of photos.[417] The facial recognition software then provides several possible "matches," and a human police officer uses those possible "matches" as investigative leads.[418]

Police use of FRT and databases in investigations is now commonplace.[419] The FBI, for example, routinely runs face recognition searches through its agency system.[420] Federal and state databases are also not limited to mug shots; now, nearly half of American adults are in a law enforcement agency's facial recognition network, through the use of state driver's license databases or other identification photos.[421]

Crucially, FRT cannot yet "match" two photographs for use as identification evidence at a trial, because of its current accuracy limitations,[422] particularly when used to identify people of color.[423] According to a 2019 NIST study evaluating the effects of race, age, and sex on facial recognition software, the

415. *See id.* at 13 n.30 ("We believe some form of state review exists in at least the following states: Arizona, Arkansas, California, Colorado, Connecticut, Florida, Illinois, Iowa, Kansas, Massachusetts, Missouri, Nebraska, New York, North Carolina, Pennsylvania, Texas, and Virginia."). *See also* FBI, *Director Comey Letter to Additional Governors on State Reviews* (June 10, 2016), https://perma.cc/ZJ3Y-3AL6.

416. Kaitlin Jackson, *Challenging Facial Recognition Software in Criminal Court,* Champion, July 2019, at 14, https://perma.cc/6HQ9-GBC6.

417. *Id.* The database can include civilian photos from the Department of Motor Vehicles.

418. *Id.*

419. *See, e.g.*, Hannah Bloch-Wehba, *Visible Policing: Technology, Transparency, and Democratic Control*, 109 Calif. L. Rev. 917, 921, 933 (2021) (noting that the extent of police use of facial recognition technology is both common and likely greater than the public is aware because of lack of transparency over its use); Claire Garvie et al., *The Perpetual Line-Up: Unregulated Police Face Recognition in America*, Geo. L. Ctr. on Priv. & Tech. (Oct. 18, 2016), https://perma.cc/SC56-9UJQ.

420. *See generally* Garvie, *supra* note 419.

421. *Id.*

422. *See, e.g.*, People v. Collins, 15 N.Y.S.3d 564, 575 (Sup. Ct. 2015) (noting that facial recognition software, like certain cutting-edge DNA techniques, "can aid an investigation, but are not considered sufficiently reliable to be admissible at a trial").

423. Use of facial recognition software has been shown to disproportionately affect Black Americans, Asian Americans, and Native Americans. *See* Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism (2018); Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (2019).

software has been least accurate in identifying Black women, even misidentifying their gender.[424] Two recent high-profile cases involved false accusations of crime against people of color.[425] In response, software datasets are growing to reflect diversity, in order to diminish error rates.[426] Still, while making datasets more representative addresses one source of error, another source of error can be the comparison of features by human examiners. Most recently, the National Academy of Sciences issued a comprehensive report on FRT, including a discussion of its uses, limitations, and recommendations for best practices.[427]

FRT output as of this writing has not yet been admitted as identification evidence at trial, and at least one trial court has observed that there is "no agreement in a relevant community of technological experts that [FRT] matches are sufficiently reliable to be used in court as identification evidence."[428] Nonetheless, courts should be aware of its investigative use and keep abreast of future developments in the field. As the 2024 NAS Report notes, judges should be ready not only to determine whether FRT as an identification technique is foundationally reliable, but also whether it is "applied reliably by an appropriately trained, competent analyst."[429] Likewise, speaker recognition technology, as of this writing, is being used to investigate crimes such as false marine distress calls, but is rarely as yet used in criminal cases as evidence.[430]

424. NIST, *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software* (Dec. 19, 2019), https://perma.cc/C5KV-HYLS.

425. *See* Kashmir Hill, *Eight Months Pregnant and Arrested After False Facial Recognition Match*, N.Y. Times, Aug. 6, 2023, https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html (Porcha Woodruff falsely accused of robbery and carjacking); Tate Ryan-Mosley, *The New Lawsuit That Shows Facial Recognition Is Officially a Civil Rights Issue*, MIT Tech. Rev., Apr. 14, 2021, https://perma.cc/89WV-E2LL (noting false theft and burglary accusation against Robert Williams in Detroit).

426. Additionally, there is a NIST OSAC Facial Identification Subcommittee that works "to develop consensus standards and guidelines for the image-based comparisons of human facial features and to provide recommendations for the research and development necessary to advance the state of the science." *See* NIST, *Facial Identification Subcommittee*, https://perma.cc/XJ4K-SPLQ.

427. Nat'l Acads. of Scis., Eng'g & Med. (NAS), Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance (2024), https://doi.org/10.17226/27397 [hereinafter 2024 NAS Report].

428. People v. Reyes, 133 N.Y.S.3d 433, 436–37 (Sup. Ct. 2020).

429. 2024 NAS Report, *supra* note 427, at 103.

430. *See generally* Peter Andrey Smith, *Can We Identify a Person from Their Voice?*, IEEE Spectrum, Apr. 15, 2023, https://perma.cc/4KUL-Q8VM (noting current investigative uses of voice recognition technology but also that it is not yet used in criminal trials in the United States).

Special Issues with Machine-Generated Feature Comparison Evidence

Feature comparison methods are being increasingly automated. Previous sections have already described some of these turns toward, or research to eventually enable, expert systems in place of human analysis in feature comparison. This section briefly addresses unique conceptual and legal issues raised by this automation.

First, a brief note about definitions. By machine-generated feature comparison evidence, we mean the conveyance by a machine of a conclusion about whether or how two features or patterns "match," such as a software program's reported "likelihood ratio" comparing the chance of seeing the evidence sample under competing hypotheses about whether the defendant is the source of the evidence. In that case, the expert system itself is generating and reporting the information output, just like a human expert might make an assertion in a report or on the witness stand. Machine-generated feature comparison is therefore distinct from mere electronic storage of, or conduits for, information. Courts have admitted machine-stored information since at least the 1970s.[431] Indeed, electronically stored information (ESI) has now been officially incorporated into the rules of civil discovery.[432] Likewise, courts routinely admit emails and social media posts made by people and communicated via computers. To be sure, such evidence raises potential authentication questions (such as whether a purported email is actually written by a particular declarant).[433] But ESI generally does not raise novel issues related to the reliability of the information itself, such as whether an assertion in an email is true.

Types of Machine-Generated Feature Comparison Evidence

The past decade has seen a precipitous rise in feature comparison techniques automated or at least facilitated by computer software. Techniques that were previously the province only of human examiners, such as firearm and toolmark

431. *See, e.g.*, United States v. Liebert, 519 F.2d 542, 543 (3d Cir. 1975) (admitting list of persons not filing tax returns, stored in IRS computers).

432. *See* Fed. R. Civ. P. 34.

433. *See* Fed. R. Evid. 902(13), (14) & advisory committee's notes to 2017 amendments. For an example of a successful authentication challenge to ESI, see *United States v. Vayner*, 769 F.3d 125 (2d Cir. 2014) (holding that admission of social media page and contents against defendant was reversible error because not properly authenticated as having been posted by him).

comparison,[434] handwriting comparison,[435] shoemark comparison,[436] latent print comparison,[437] facial recognition,[438] and arson analysis based on burn or debris patterns,[439] can now potentially be done by algorithm. The Government Accounting Office issued a report in 2021 on issues raised by automated forensic methods—in particular DNA, latent print analysis, and facial recognition.[440] This reference guide focuses on non-DNA comparison disciplines. Nonetheless, because existing cases addressing forensics algorithms are all in the DNA context, judges should have a general understanding of such cases. Competing software programs now purport to interpret complex DNA mixtures, determine whether a defendant's profile is consistent with the mixture, and report associated match statistics. Courts have generally admitted the results of such programs—typically, but not exclusively, offered by the prosecution[441]—over *Frye/Daubert* objections.[442] The few exceptions have been cases in which a local laboratory did not perform internal validation studies before using the software,[443] where a mixture had a particularly

434. *See generally,* Eric Hare, Heike Hofmann & Alicia Carriquiry, *Automatic Matching of Bullet Land Impressions*, 11 Annals Applied Stat. 2332 (2017), https://doi.org/10.1214/17-AOAS1080 (creating an open-source algorithm for automatically comparing certain striations, after removing class characteristics, on fired ammunition).

435. *See, e.g.*, Amy M. Crawford, Danica M. Ommen & Alicia L. Carriquiry, *A Statistical Approach to Aid Examiners in the Forensic Analysis of Handwriting*, 68 Annals Applied Stat. 1768 (2023), https://doi.org/10.1111/1556-4029.15337.

436. Soyoung Park & Alicia Carriquiry, *An Algorithm To Compare Two-Dimensional Footwear Outsole Images Using Maximum Cliques and Speeded-Up Robust Feature*, 13 Stat. Analysis & Data Mining: ASA Data Sci. J. 188 (2020), https://doi.org/10.1002/sam.11449.

437. *See* Simon A. Cole et al., *Beyond the Individuality of Fingerprints: A Measure of Simulated Computer Latent Print Source Attribution Accuracy*, 7 L. Prob. & Risk 165, 166 (2008), https://doi.org/10.1093/lpr/mgn004 (explaining how the AFIS database returns several "candidate" prints to the analyst).

438. *See, e.g.*, Lynch v. State, 260 So.3d 1166 (Fla. Dist. Ct. App. 2018).

439. *See, e.g.*, Sander Korver et al., *Artificial Intelligence and Thermodynamics Help Solving Arson Cases*, 10 Sci. Reps. 20502 (2020), https://www.nature.com/articles/s41598-020-77516-x.

440. *See* Karen L. Howard, *Forensic Technology: Algorithms Offer Benefits for Criminal Investigations, but a Range of Factors Can Affect Outcomes*, Gov't Accountability Off. (Jan. 24, 2024), https://perma.cc/SW4Q-TX3Z.

441. One such program, TrueAllele, has been used by a handful of defendants to demonstrate their innocence. *See generally, TrueAllele Helps Free Innocent Indiana Man After 24 Years in Prison*, Cybergenetics, Apr. 26, 2016, https://perma.cc/RA7Q-XV92.

442. *See* John S. Hausman, *Lost Shoe Led to Landmark DNA Ruling—And Now, Nation's 1st Guilty Verdict*, MLive, https://perma.cc/UA3W-F8XF (reporting a conviction in Michigan as the first in the United States to be based in part on the STRmix software after the defense contested its admissibility); *Trials*, Cybergenetics, https://perma.cc/V9ZK-MFP6 (listing over fifty cases in which TrueAllele has been admitted).

443. *See* Decision & Order on DNA Analysis Admissibility, People v. Hillary, Indictment No. 2015-15 (N.Y. Cnty. Ct. Aug. 26, 2016), http://perma.cc/3TFA-8VA5 (excluding STRmix results under *Frye*).

high number of contributors,[444] where the evidence involved a minor contributor to a mixture,[445] or where the software was discontinued.[446] In October 2019, in *United States v. Gissantaner*, a federal district judge ruled the results of one program inadmissible in a case involving "low copy number" DNA,[447] but the Sixth Circuit reversed the following spring.[448]

One particular 2016 case, although involving DNA, is important to be aware of, as it will likely be invoked in future challenges to results of other automated forensic analyses. In a high-profile New York homicide case, the two main DNA programs—STRmix™ and TrueAllele®—reached very different conclusions. The case involved the strangulation of a young boy. Police suspicion fell upon Oral Hillary, a former college soccer coach who had dated the boy's mother.[449] Another former boyfriend, a deputy sheriff who had been physically violent with the mother, was cleared of suspicion. No physical evidence linked either man to the scene. But analysts eventually collected a DNA mixture, too complex to be analyzed by humans, from under the boy's fingernail. Police in 2013 gave the DNA fingernail data to the proprietor of TrueAllele. In 2014, TrueAllele reported "no statistical support for a match" with Hillary.[450] Indeed the TrueAllele CEO would later insist the evidence suggested Hillary was excluded as a contributor, based on the reported likelihood ratio.[451] A year later, a new district attorney had the DNA data analyzed through STRmix, which reported that Hillary was 300,000 times more likely than a random person to have contributed to the mixture.[452] A trial judge excluded the STRmix results because of a lack of internal validation,[453] and

444. *See* Order Excluding DNA Evidence, United States v. Alfonzo Williams, Case No. 3:13-cr-00764-WHO-1 (N.D. Cal. Apr. 29, 2019) (Orrick, J.) (excluding results of "Bulletproof" genotyping software in case with potentially more than four contributors).

445. *See* United States v. Lewis, 442 F. Supp. 3d 1122 (D. Minn. 2020).

446. *See* Shayna Jacobs, *Judge Tosses Out Two Types of DNA Evidence Used Regularly in Criminal Cases*, N.Y. Daily News, Jan. 5, 2015, https://perma.cc/A9ZL-BMJ6 (reporting a Brooklyn judge excluded results from low copy number DNA testing and Forensic Statistical Tool testing).

447. United States v. Gissantaner, 417 F. Supp. 3d 857 (W.D. Mich. 2019) (Neff, J.), *rev'd*, 990 F.3d 457 (6th Cir. 2021).

448. *Gissantaner*, 990 F.3d 457. Rehearing en banc was denied, *id.,* and the case appears to have since ended in a change of plea. *See* https://perma.cc/E765-9FW6.

449. *See, e.g.*, Jesse McKinley, *Tensions Simmer over Race as Town Reels from Boy's Killing*, N.Y. Times, Mar. 5, 2016, at A1, https://perma.cc/AE3L-NPVW.

450. W.T. Eckert, *Hillary Trial Slated for Aug. 1*, Watertown Daily Times (Mar. 2, 2016).

451. *See* John Buckleton, Memorandum, *People v. Hillary* (2017), https://perma.cc/H3UE-GT7U (proprietor of STRmix explaining the discrepancies between STRmix and TrueAllele likelihood ratios in the *Hillary* case).

452. Notice of Motion to Preclude at 10, People v. Hillary, Indictment No. 2015-15 (N.Y. Cnty. Ct. May 31, 2016), https://perma.cc/2H69-DJJT.

453. *See* Decision & Order, People v. Hillary, Indictment No. 2015-15 (N.Y. Cnty. Ct. Aug. 26, 2016) (Catena, J.).

Hillary was acquitted.[454] The creator of STRmix has since published a memorandum online, explaining how STRmix's approach to the evidence in *Hillary* is different from TrueAllele's approach.[455] The case is an example of two programs, both deemed reliable by courts under *Daubert* or *Frye*, coming to significantly different conclusions based on the same forensic data.[456] Judges should also be aware of a 2021 NIST study exploring the reliability of probabilistic genotyping software, given that its findings and recommendations might be relevant to evaluation of other automated methods.[457]

Recurring Legal Issues

This section offers an overview of legal issues potentially raised by machine-generated feature comparison evidence,[458] some of which are already the subject of federal litigation.

Rule 702/Reliability Issues

If a computer program or machine output is the method upon which a human feature comparison expert's opinion is based, then the algorithm might be subject to *Daubert* scrutiny.[459]

454. Jesse McKinley, *Race, Jilted Love and Acquittal in Boy's Killing*, N.Y. Times, N.Y. Ed., at A1 (Sept. 29, 2016).

455. *See* Buckleton, *supra* note 451.

456. The study emphasized the importance, in determining software accuracy in a given case, of validation studies covering the particular "factor space" that a case falls into (such as, in the DNA mixture context, the quantity of DNA, number of contributors, etc.). *See* John M. Butler et al., NIST, DNA Mixture Interpretation: A NIST Scientific Foundation Review (2021), https://doi.org/10.6028/NIST.IR.8351-draft (noting that report is in draft form but that comment period has closed).

457. *Id.*

458. For a general discussion of legal issues related to algorithmic feature comparison evidence, see generally Andrea Roth, *The Use of Algorithms in Criminal Adjudication*, *in* Cambridge Handbook on the Law of Algorithms 407 (Woodrow Barfield ed., 2020), https://doi.org/10.1017/9781108680844 (describing the rise of algorithmic proof in criminal cases, and the legal and ethical issues raised); Edward K. Cheng & G. Alexander Nunn, *Beyond the Witness: Bringing a Process Perspective to Modern Evidence Law*, 97 Tex. L. Rev. 1077 (2019) (discussing how evidence law and Confrontation Clause jurisprudence should accommodate computer-generated evidence); Andrea Roth, *Machine Testimony*, 126 Yale L.J. 1972 (2017) (describing the rise of machine-generated conveyances of information as proof, and testimonial safeguards that might be deployed to better regulate such proof in a manner analogous to human assertions).

459. *See* Fed. R. Evid. 702 (applying only to expert "witnesses" who testify). *Cf.* People v. Lopez, 286 P.3d 469, 478 (Cal. 2012) (portion of spectrometer readings offered to prove blood-alcohol concentration, without additional sworn testimony or certification by human expert, not

Reliability concerns have arisen with respect to several types of machine-generated conclusions both within and outside the feature comparison context, including driving estimates,[460] blood alcohol testing,[461] translation of the federal sentencing guidelines into computer code,[462] Apple's "Find My iPhone" tracking when used in theft and robbery cases,[463] "minor miscode[s]" in DNA software,[464] and inaccurate allelic frequencies used by software to generate DNA match statistics.[465] Additionally, as discussed in the section titled "Facial Recognition Software and Other Emerging Methods" above, machine learning algorithms can misidentify or misclassify people or events because of limitations in the training dataset or other analytical problems.[466] Machines learn how to characterize new data by training on data already categorized by a person or the machine itself. If the dataset is too small or too unrepresentative of future items to be classified, the algorithm might infer a pattern or linkage in the training data that does not actually mirror real life (overfitting) or might try to account for too many variables, rendering the training data inadequate for learning (the so-called curse of dimensionality).[467]

With respect to issues that might arise in a *Daubert* hearing related to software results, some commentators have argued that validation studies alone might be insufficient to scrutinize certain algorithmic proof. For example, while validation studies might show that a certain interpretive software program boasts a

subject to the Confrontation Clause because not the statement of a human expert, over dissent by Justice Liu).

460. *See, e.g.*, Jianniney v. State, 962 A.2d 229, 232 (Del. 2008) (excluding Mapquest driving estimates as inadmissible hearsay and noting reliability concerns).

461. *See* Andrea Roth, *Trial by Machine*, 104 Geo. L.J. 1245, 1271–72 (2016) (citing sources with respect to problems with Intoxilyzer 8000 and 5000 readings, due to programming errors); Roth, *Machine Testimony*, *supra* note 458, at 1995 (discussing litigation over code errors with the Alcotest 7110).

462. *See* Steven R. Lindemann, Commentary, *Published Resources on Federal Sentencing*, 3 Fed. Sent'g Rep. 45, 45–46 (1990), https://doi.org/10.2307/20639272.

463. *See* Lawrence Mower, *If You Lose Your Cellphone, Don't Blame Wayne Dobson*, Las Vegas Rev.-J., Jan. 13, 2013, https://perma.cc/2G7G-89QX.

464. Roth, *Trial by Machine*, *supra* note 461, at 1276.

465. *See Notice of Amendment of the FBI's STR Population Data Published in 1999 and 2001*, FBI (2015), https://perma.cc/TDF7-WMLA.

466. *See also Face Recognition*, Elec. Frontier Found., https://perma.cc/W648-PQWF (citing Brendan F. Klare et al., *Face Recognition Performance: Role of Demographic Information*, 7 IEEE Transactions Info. Forensics & Sec. 1789 (2012), https://doi.org/10.1109/TIFS.2012.2214212 (showing higher false positive rates for African Americans)).

467. *See generally* The Discipline of Organizing: Informatics Edition (Robert J. Glushko ed., 4th ed. 2013); Harry Surden, *Machine Learning and Law*, 89 Wash. L. Rev. 87 (2014); Pedro Domingos, The Master Algorithm: How the Quest for the Ultimate Learning Machines Will Remake Our World (2015); Alice Zheng, Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls (2015).

low false positive rate, such studies cannot so easily determine whether a reported likelihood ratio is inaccurate:

> Laboratory procedures to measure a physical quantity such as a concentration can be validated by showing that the measured concentration consistently lies with an acceptable range of error relative to the true concentration. Such validation is infeasible for software aimed at computing a[] [likelihood ratio] because it has no underlying true value (no equivalent to a true concentration exists). The [likelihood ratio] expresses our uncertainty about an unknown event and depends on modeling assumptions that cannot be precisely verified in the context of noisy [crime scene profile] data.[468]

In sum, judges faced with determining foundational validity and validity as applied of machine-generated feature comparison conclusions will need to consider separately whether the software's classifications are reliable (e.g., identification versus nonidentification; consistent versus inconsistent) and whether the software's reported "scores" are reliable (such as a match statistic or score where the ground truth is not known). Judges will face the same questions as with other feature comparison methods in terms of what counts as a well-designed validation study, and how many well-designed studies are necessary, to determine a program's error rate. They will also need to decide whether some issues (such as the reliability of reported scores that cannot be tested through traditional black box validation studies) require greater access to the source code underlying the software, a discovery issue discussed below.

Hearsay and Confrontation

Some litigants have tried to argue, most unsuccessfully, that a machine-generated conclusion offered as evidence is "hearsay" and therefore inadmissible unless it comes within an exception to the rule against hearsay. Hearsay, under Federal Rules of Evidence 801 and 802, is an out-of-court "assertion" by a "person," "intended" as an assertion, and offered to prove the truth of the matter asserted.[469] By the explicit language of these rules, as numerous courts have now held,[470] a factual claim rendered by a machine rather than a person cannot be

468. Christopher D. Steele & David J. Balding, *Statistical Evaluation of Forensic DNA Profile Evidence*, 1 Ann. Rev. Stat. & Its Application 361, 380 (2014), https://doi.org/10.1146/annurev-statistics-022513-115602.

469. *See, e.g.*, Fed. R. Evid. 801, 802 (defining hearsay as an out-of-court "statement," "intended" as an assertion, offered to prove the truth of the matter asserted, and stating that hearsay is presumptively inadmissible).

470. *See, e.g.*, People v. Lopez, 286 P.3d 469, 478 (Cal. 2012) (holding that gas chromatograph "raw" data is not hearsay); United States v. Lizarraga-Tirado, 789 F.3d 1107, 1109 (9th Cir. 2015) (holding that Google Earth output is not hearsay).

hearsay. To be sure, some courts used to treat computer-generated results as hearsay and admit them under the "business records" hearsay exception for records created in the regular course of business.[471] But commentators were critical of this practice, pointing out that these courts were conflating electronically stored records, inputted by humans but stored by computers, with electronically generated records, where the computer itself creates and reports information.[472]

Some courts and commentators have alternatively suggested that an algorithm's conclusions might be the hearsay statements of the algorithm's creator, such as a computer programmer, and that machine-generated proof is admissible so long as the programmer is subject to cross-examination.[473] Others have argued against this view of algorithmic evidence, noting that the programmer has not uttered the resulting information, and might not even fully understand the thousands of steps taken by the program to generate the information, particularly in more advanced forms of artificial intelligence (AI).[474]

Some litigants and commentators have further argued that admission of machine-generated conclusions without sufficient opportunity for adversarial scrutiny may violate the Sixth Amendment's Confrontation Clause,[475] which guarantees an accused the right "to be confronted with the witnesses against him." To be sure, under existing U.S. Supreme Court precedent, the right of confrontation applies only to "testimonial hearsay,"[476] which presumably only covers certain human assertions. While the Court has not squarely addressed the question, Justice Sotomayor has suggested in a concurring opinion that "raw data" from a machine might not be "testimonial."[477] Not surprisingly, then,

471. Apparently, this approach was buttressed by an oft-cited 1974 article arguing that "[c]omputer generated evidence will inevitably be hearsay." Jerome J. Roberts, *A Practitioner's Primer on Computer-Generated Evidence*, 41 U. Chi. L. Rev. 254, 272 (1974).

472. *See generally* Adam Wolfson, Note, *"Electronic Fingerprints": Doing Away with the Conception of Computer-Generated Records as Hearsay*, 104 Mich. L. Rev. 151 (2005) (noting, and criticizing, courts' treatment of computer-generated business records as hearsay).

473. *See, e.g.,* People v. Wakefield, 38 N.Y.3d 367, 385–86 (2022) (holding that admission of TrueAllele results did not violate the Confrontation Clause so long as the creator, Mark Perlin, was on the witness stand); United States v. Washington, 498 F.3d 225, 229 (4th Cir. 2007) (explaining defendant's argument that the "raw data" of a chromatograph was the "hearsay" of the "technicians" who tested the defendant's blood sample for PCP and alcohol using the machine); Karen Neville, *Programmers and Forensic Analyses: Accusers Under the Confrontation Clause*, 10 Duke L. & Tech. Rev. 1, 8–9 (2011) (arguing that "the programmer" is "the 'true accuser'—not the machine merely following the protocols he created").

474. *See, e.g.*, Roth, *Machine Testimony*, *supra* note 458, at 1986. *See generally* Cheng & Nunn, *supra* note 458 (describing potential inaccuracies of "process-based" proof and suggesting enhanced pretrial discovery, not cross-examination of an individual expert, as the most appropriate safeguard).

475. *See, e.g.*, *Wakefield*, 38 N.Y.3d at 385 (arguing that admission of TrueAllele results without disclosure of the source code violated the Confrontation Clause).

476. Crawford v. Washington, 541 U.S. 36 (2004).

477. Bullcoming v. New Mexico, 564 U.S. 647, 674 (2011) (Sotomayor, J., concurring).

most courts have thus far rejected arguments that the admission of a machine's claim without disclosing certain information about a machine's processes (such as source code) violates the Confrontation Clause.[478] Still, several commentators have argued that any approach categorically exempting machine assertions from the right of confrontation may soon become untenable, especially as machine-generated conclusions become ever more complex, opaque, and interpretive, akin to human expert testimony.[479]

Discovery Issues

A final recurring legal issue related to machine-generated feature comparison evidence is disputes over pretrial requests for disclosure of the source code or for access to a research license to conduct software audits. The dispute stems from the fact that many (though not all) of the programs underlying machine-generated proof are proprietary,[480] and program developers often argue, in

478. *See, e.g.*, *Wakefield*, 38 N.Y.3d at 385 (2022) (holding that admission of TrueAllele results without disclosure of the source code did not violate the Confrontation Clause); People v. Lopez, 286 P.3d 469, 478 (Cal. 2012) (holding that gas chromatograph "raw" data is not hearsay and thus its admission did not implicate the Confrontation Clause). *But see* People v. Chubbs, No. B258569, 2015 WL 139069, at *4 (Cal. Ct. App. Jan. 9, 2015) (noting that the trial court had invoked the Confrontation Clause in ordering disclosure of source code to facilitate cross-examination of programmer); Order on Procedural History and Case Status in Advance of May 25, 2016 Hearing, United States v. Michaud, No. 3:15-CR-05351RJB, 2016 WL 337263 (W.D. Wash. Jan. 28, 2016) (noting due process right to examine source code of government's Network Investigative Technique (NIT) used to hack defendant's computer).

479. *See* Andrea Roth, *What Machines Can Teach Us About "Confrontation,"* 60 Duquesne L. Rev. 210 (2022) (explaining that a view of confrontation as merely guaranteeing cross-examination at trial is ahistorical); Cheng & Nunn, *supra* note 458 (arguing that machine-generated proof cannot be exempt from confrontation); Roth, *Machine Testimony*, *supra* note 458 (same); Erin Murphy, *The Mismatch Between Twenty-First-Century Forensic Evidence and Our Antiquated Criminal Justice System*, 87 S. Calif. L. Rev. 633, 657–58 (2014); David Alan Sklansky, *Hearsay's Last Hurrah*, 2009 Sup. Ct. Rev. 1, 67 (2009) (urging a view of confrontation as "a meaningful opportunity to test and to challenge the prosecution's evidence"); *see also id*. at 7 (quoting Daniel H. Pollitt, *The Right of Confrontation: Its History and Modern Dress*, 8 J. Pub. L. 381, 402 (1959)) ("The Confrontation Clause could be read broadly to guarantee criminal defendants a meaningful opportunity to challenge—'to know, to examine, to explain, and to rebut'—the proof offered against them.").

480. *See, e.g.*, State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (denying litigants access to source code of actuarial tool used in parole hearing); *Chubbs*, 2015 WL 139069, at *8–9 (noting that DNA mixture interpretation software TrueAllele is proprietary); Order, United States v. Michaud, No. 3:15-CR-05351RJB (W. Dist. Wash. May 18, 2016) (noting that government's malware used in child pornography investigation was proprietary); Luke Broadwater & Scott Calvert, *City in $2 Million Dispute With Xerox Over Camera Tickets*, Balt. Sun, Apr. 24, 2013, https://perma.cc/T2X7-F2PS (noting that Xerox refused to disclose source code for red light camera system). *But see* Hare, Hofmann & Carriquiry, *supra* note 434 (creating an open-source algorithm for automatically comparing certain striations, after removing class characteristics, on fired ammunition, and noting the benefits of open-source software for transparency and improvement).

response to discovery requests for information like source code, that such information is protected by a "trade secrets privilege."[481] Some have suggested that a trade secret privilege might be necessary to incentivize development of high-quality algorithms for use in criminal justice,[482] while others have argued that such a privilege is inappropriate and unnecessary.[483] In particular, Professor Rebecca Wexler has argued that the emergence of a trade secrets privilege in criminal cases was an historical accident, and that substantive trade secret doctrine offers software developers sufficient protection to incentivize innovation. Until recently, courts appeared to have largely accepted proprietors' claims of privilege.[484] At least two courts have granted defense requests for source code in cases involving DNA software, however.[485] Meanwhile, U.S. Representative Mark Takano (D-CA) introduced a bill (originally introduced in 2019) titled "Justice in Forensic Algorithms Act of 2024" that would require disclosure of source code and remove any trade secret privilege over such code.[486]

Admission of machine-generated feature comparison conclusions might raise other discovery issues as well. For example, a party might ask for pretrial access to a program to manipulate its inputs, in the same way that a party might seek to depose or interview an opposing party's expert witness before trial.[487] Next, parties might ask for access to the prior statements or runs of software programs in a given case. While there is no Jencks Act for machine conveyances, judges will have to decide whether to grant access to such statements as fairness demands.[488] In addition, judges might face requests by a party for access to the

481. *See, e.g.*, Roth, *Machine Testimony*, *supra* note 458, at 2028 (discussing trade secrets); Christian Chessman, Note, *A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution*, 105 Calif. L. Rev. 179, 212 (2017) (discussing the trade secret privilege in criminal cases with respect to source code).

482. *See, e.g.*, Edward J. Imwinkelried, *Computer Source Code: A Source of the Growing Controversy over the Reliability of Automated Forensic Techniques*, 66 DePaul L. Rev. 97 (2016).

483. *See* Chessman, *supra* note 481, at 212; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343 (2018); Jason Tashea, *Trade Secret Privilege Is Bad for Criminal Justice*, A.B.A. J., July 30, 2019, https://perma.cc/CKY9-HLVN.

484. Wexler, *supra* note 483, at 1352–53.

485. The New Jersey case is *State v. Pickett*, 246 A.3d 279 (N.J. Super. Ct. App. Div. 2021). The state has since withdrawn its intent to introduce the TrueAllele results, and thus this litigation appears to be moot. The other case is *United States v. Ellis*, No. 19-369, 2021 WL 1600711 (W.D. Pa. Apr. 23, 2021). As of this writing, the parties had agreed upon terms of a protective order and review by experts had just begun.

486. *See* H.R. 7394, 118th Cong. (Feb. 15, 2024), https://www.congress.gov/bill/118th-congress/house-bill/7394/committees.

487. *Cf.* Jennifer L. Mnookin, *Repeat Play Evidence: Jack Weinstein, "Pedagogical Devices," Technology, and Evidence*, 64 DePaul L. Rev. 571, 588 (2015) (arguing that demonstrative evidence in the form of complex algorithms should have this as a condition of admissibility).

488. *See, e.g.*, Roth, *Machine Testimony*, *supra* note 458, at nn.411–19 & accompanying text (explaining why the *Jencks* case (Jencks v. United States, 353 U.S. 657 (1957)) can and should apply to machines and why it would be consistent with the Jencks Act).

data on which a machine-learning algorithm was "trained" to classify future objects or events. A party might seek this dataset to argue that the algorithm was trained on data that was not representative of the person, object, or event being classified in the case (such as facial recognition software claiming to have identified an African American as having been at a crime scene, where the software was trained largely on white faces). The proponent of the evidence might argue that the data are covered by privacy laws and cannot be disclosed. Just as judges have always had to grapple with access to privacy-protected data in contexts like a victim's medical records, judges will need to consult applicable data-protection laws, with the accused's constitutional right to present a defense in mind, in resolving these issues.[489]

489. For a brief discussion of intellectual property and data privacy issues related to machine learning training data, see Brittany Bacon et al., *Training a Machine Learning Model Using Customer Proprietary Data: Navigating Key IP and Data Protection Considerations*, *in* Pratt's Privacy and Cybersecurity Law Report, LexisNexis 233 (Steven A. Meyerowitz et al. eds., Oct. 2020).

Glossary of Terms

This glossary has been adapted from a general literature review related to the topics included in this reference guide. An ongoing project by NIST's Organization of Scientific Area Committees (OSAC) is also creating a list of preferred terms for recurring concepts in forensic science.[490]

ACE-V. Acronym for the method many feature comparison analysts use and refer to as Analysis, Comparison, Evaluation, and Verification.

AFIS. Acronym for Automated Fingerprint Identification System used by law enforcement to rapidly screen large numbers of prints in a national database, which are then evaluated by analysts. While AFIS is useful for screening, the system is not currently able to determine that a particular pair of prints has a common source.

algorithm. A list of steps capable of being performed by a machine.

"black box" study. A validation study designed to determine the extent to which a forensic examiner applying a feature comparison method reaches the correct conclusion (where the ground truth is known), without having to know anything about how the method itself—which may well remain a "black box"—works.

class characteristics. Characteristics believed to be shared by a group of persons or objects (e.g., ABO blood types).

contextual bias. Bias that occurs when a forensic examiner's judgment while engaged in feature comparison is influenced by irrelevant information about the facts of a case.

double-"blind" or double-anonymized research or testing. Research or testing in which the respondent and the interviewer are not given information that will alert them to the anticipated or preferred pattern of response. While the phrases "blind" and "double-blind" have been eliminated by many users over concerns that such language is ableist, judges will still frequently encounter these terms in literature discussing feature comparison disciplines.

false negative rate. The rate at which an examiner, using a feature comparison method, erroneously concludes that two features or patterns are inconsistent when, in fact, they are consistent. Note that there is a debate in the literature regarding how to incorporate an examiner's determination of "inconclusive" (in a testing context, not in casework), when the ground truth is that two features or patterns are consistent, into a false negative rate.

490. *See* NIST, *OSAC Lexicon,* https://perma.cc/VKY4-KEND.

false positive rate. The rate at which an examiner, using a feature comparison method, erroneously concludes that two features or patterns are consistent when, in fact, they are not consistent. Note that there is a debate in the literature regarding how to incorporate an examiner's determination of "inconclusive" (in a testing context, not in casework), when the ground truth is that two features or patterns are inconsistent, into a false positive rate.

FDE. Acronym for Forensic Document Examiners, who compare known and unknown writing samples and perform various tests on documents, such as whether a particular ink formulation existed on the purported date of a writing. Also referred to as Questioned Document Examiners (QDE).

fingerprint conclusions. A fingerprint examiner's conclusion as to the strength of the evidence in supporting an inference about whether two impressions came from the same source. Some laboratories, such as the Department of Justice, guide examiners to use one of three conclusions: source identification, a source exclusion, or inconclusive determination.

fingerprint inconclusive determination. The label some examiners use to describe their determination that there is "insufficient quantity and/or clarity" between the two impressions for the examiner to arrive at a source identification or exclusion of fingerprints.

fingerprint source exclusion. The label some examiners use to describe their determination that the two sets of fingerprints did not come from the same source.

fingerprint source identification. A label some examiners use to describe their determination that two sets of fingerprint prints—known and unknown—came from the same person. Other examiners might eschew this "identification" language and instead make statements about the strength of the evidence, such as that there is "extremely strong support" that the impressions came from the same source and "extremely weak support" that they came from different sources.

individual characteristic. A characteristic believed to be unique to an object or person.

individualization. The claim that a feature comparison method can scientifically determine that a feature or pattern (such as a fingerprint, or toolmark, or bitemark) was produced by one source, to the exclusion of all other sources. This claim is notably different from the concept of "uniqueness," a claim that a particular biological trait (such as the ridges on one's fingers) is unique to each person.

latent prints. Fingerprints found at a crime scene or in another location, which may vary in quality or number. Latent prints are compared to record prints to determine whether there is consistency between the two.

likelihood ratio (LR). A ratio in which the numerator is the chance of seeing a particular type of evidence (such as a DNA "match") given one hypothesis (e.g., that the defendant is the source of the DNA) and the denominator is the chance of seeing the evidence (e.g., the DNA "match") given the alternative hypothesis (such as that the defendant is *not* the source of the DNA). An LR is part of Bayes' Theorem, a means of updating a prior guess about the probability of an event with new evidence (incorporated into the LR) to arrive at a new guess ("posterior probability"). Computer programs purporting to interpret complex DNA mixtures typically report their conclusions in the form of LRs.

machine learning. An algorithm in which the computer is "trained" on a dataset (such as emails already classified as "spam" or "not spam") such that the computer "learns" to classify objects in the future (such as determining whether future emails are "spam" versus "not spam"). More broadly, machine learning refers to any system in which a machine learns from experience and improves its performance on an assigned task over time.

OSAC. The Organization of Scientific Area Committees, an organization under the direction of the National Institute of Standards and Technology (NIST), under the United States Department of Commerce. OSAC focuses on developing and approving standards to govern forensic science service providers.

PCAST. President's Council of Advisors on Science and Technology.

QDE. See FDE.

record prints. Fingerprints that are rolled onto a fingerprint card or digitized and scanned into a file. These prints are the ones stored in databases and against which unknown prints are compared.

reliability. The extent to which the same results are obtained in each instance in which a test or method is performed—its consistency. Reliability can be further divided into repeatability, the ability of one examiner to repeat the same test results using the same method under the same conditions; and reproducibility, the ability of another examiner to repeat the same test results using the same method under the same conditions.

source code. The code written by a computer programmer, in human-readable form, to a computer containing the instructions for what to do.

SWGDOC. The Scientific Working Group for Forensic Document Examination.

SWGFAST. The Scientific Working Group on Friction Ridge Analysis, Study, and Technology.

toolmark conclusions. A toolmark examiner's conclusion as to the strength of the evidence in supporting an inference about whether two toolmarks came from the same source. Some laboratories, such as the Department of Justice,

guide examiners to use one of three conclusions: source identification, source exclusion, or inconclusive determination. For a description of each, see Fingerprint Conclusions.

ULTR. The Department of Justice's guidelines on Uniform Language for Testimony and Reports.

validity. The ability of a test to measure what it is supposed to measure—its accuracy. Validity includes reliability, but the converse is not necessarily true.

Reference Guide on Human DNA Identification Evidence

DAVID H. KAYE

David H. Kaye, M.A., J.D., is Regents Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, and Distinguished Professor of Law and Academy Professor Emeritus, Pennsylvania State University School of Law.

Author's Note: Research for this reference guide was completed in 2023. I am grateful not only to the National Academies of Sciences, Engineering, and Medicine and Federal Judicial Center reviewers and staff, but also to Bruce Budowle, John Butler, Michael Coble, David Housman, Jarrah Kennedy, Swathi Kumar, and Bruce Weir for their comments on portions of a draft of this reference guide.

CONTENTS

FIGURES

TABLE

Introduction

Deoxyribonucleic acid, or DNA, is a molecule that encodes the genetic information in all living organisms. Its chemical structure was elucidated in 1953. More than 30 years later, samples of human DNA began to be used in the criminal justice system, primarily in cases of rape or murder. The evidence has been the subject of extensive scrutiny by lawyers, judges, and the scientific community. DNA evidence is admissible in all jurisdictions, but there are many types of forensic DNA analysis, and still more are being developed. Questions of admissibility and weight of DNA evidence arise as advancing methods of biochemical and statistical analysis, along with novel applications of established methods, are introduced. Moreover, questions about the appropriate balance between law-enforcement goals and individual privacy and security also emerge when law-enforcement officials collect and analyze human DNA samples. This reference guide addresses technical issues that are important when considering the admissibility of and weight to be accorded analyses of human DNA and the related molecule RNA (ribonucleic acid) in trials. In addition, this guide describes ways in which human DNA assists in criminal investigations, and it identifies a number of the legal and policy questions raised by investigative genetics.

A Brief History of DNA Evidence

"DNA evidence" refers to the results of chemical or physical tests that directly reveal differences in DNA molecules found in organisms as diverse as bacteria, plants, and animals.[1] The technology for establishing the identity of individuals from the DNA in blood, semen, and other bodily fluids became available to law-enforcement agencies in the mid to late 1980s.[2] The judicial reception of DNA evidence can be divided into at least six phases.[3] The first phase was one of rapid acceptance. Initial praise for RFLP (restriction fragment length polymorphism)

1. Differences in DNA also can be revealed by differences in the proteins that are made according to the "instructions" in a DNA molecule. Blood-group factors, serum enzymes and proteins, and tissue types all reveal information about the DNA that codes for these chemical structures. Such immunogenetic testing, including the application of antibodies to detect cell-surface antigens, predates the more direct DNA testing that is the subject of this guide. On the nature and admissibility of such testing, see, for example, David H. Kaye, *The Double Helix and the Law of Evidence* 5–19 (2010); 1 *McCormick on Evidence* § 205(B) (Robert Mosteller ed., 8th ed. 2020). *See generally* Liesa L. Richter & Daniel J. Capra, *Reference Guide on the Admissibility of Scientific Evidence*, in this manual.

2. The first reported appellate opinion is *Andrews v. State*, 533 So. 2d 841 (Fla. Dist. Ct. App. 1988).

3. The description that follows is adapted and updated from 1 *McCormick on Evidence*, *supra* note 1, § 205(B). Beyond these six phases for the principal methods of human DNA profiling encountered in courts, the supplementation of STR profiling with other genetic markers for identification related to single-nucleotide differences is beginning.

testing in homicide, rape, paternity, and other cases was effusive.[4] Expert testimony rarely was countered, and courts readily admitted DNA evidence.

In a second wave of cases, however, defendants pointed to problems at two levels—controlling the experimental conditions of the analysis and interpreting the results. Concerted attacks by defense experts with impressive credentials led to the rejection of specific proffers on the grounds that the testing was not sufficiently rigorous.[5]

A different attack on DNA profiling that began in cases during this period led to a third wave of cases in which many courts held that estimates of the probability of a coincidentally matching DNA profile were inadmissible. These estimates relied on a simple population-genetics model for the frequencies of DNA profiles, and some prominent scientists claimed that the applicability of the mathematical model had not been adequately verified. A heated debate on this point spilled over from courthouses to scientific journals and convinced the supreme courts of several states that general acceptance was lacking. A 1992 report of the National Academy of Sciences (NAS) proposed a more "conservative" computational method as a compromise,[6] and this seemed to undermine the claim of scientific acceptance of the procedure that was in general use.

An outpouring of critiques of the report led to the formation of a second NAS committee. This panel concluded in 1996 that the usual method of estimating frequencies in broad population groups generally was sound, and it proposed improvements and additional procedures for estimating frequencies in subgroups.[7] In the corresponding fourth phase of judicial scrutiny of DNA evidence, the courts almost invariably returned to the earlier view that the statistics associated with DNA profiling are generally accepted and scientifically valid.

4. *E.g.*, People v. Wesley, 140 Misc. 2d 306, 533 N.Y.S.2d 643, 644 (Alb. Cnty. Ct. 1988) ("the single greatest advance in the 'search for truth' . . . since the advent of cross-examination"). DNA testing also quickly became powerful evidence of kinship in child support, immigration, and other cases involving genetic parentage or other relationships within a family.

5. Moreover, a minority of courts, perhaps concerned that DNA evidence might be conclusive in the minds of jurors, added a "third prong" to the general-acceptance standard of *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). This augmented *Frye* test requires not only proof of the general acceptance of the ability of science to produce the type of results offered in court, but also of the proper application of an approved method on the particular occasion. For commentary, see David L. Faigman et al., *Modern Scientific Evidence* § 30:9 n.5 (2022–2023 ed.) (citing articles); David H. Kaye et al., *The New Wigmore, A Treatise on Evidence: Expert Evidence* § 7.3.3(a)(2) (3d ed. 2021) (criticizing the approach); Joseph G. Petrosinelli, Comment, *The Admissibility of DNA Typing: A New Methodology*, 79 Geo. L. J. 313, 327–31 (1990) (similar criticism); *cf.* Ming W. Chin et al., *Forensic DNA Evidence: Science and the Law* § 11:6 (2022) (describing California's "limited third-prong hearing").

6. National Research Council, DNA Technology in Forensic Science (1992) [hereinafter NRC I].

7. National Research Council, The Evaluation of Forensic DNA Evidence (1996) [hereinafter NRC II].

In the fifth phase of the judicial evaluation of DNA evidence, results obtained with PCR-based methods[8] entered the courtroom. The opinions were practically unanimous in holding that the PCR-based procedures rested on a solid scientific foundation and were generally accepted in the scientific community. Before long, forensic scientists settled on the use primarily of one type of DNA variation—known as short tandem repeats, or STRs—to include or exclude individuals as the source of crime-scene DNA. Investigative databases of STR profiles from men, women, and children convicted (and sometimes only arrested for) specified crimes grew by leaps and bounds.[9] Matching these recorded profiles to STR profiles of crime-scene samples generated seemingly magical investigative leads.

The sixth phase of judicial scrutiny is still in progress. With the extension of STR profiling to minute quantities of DNA (low template, or LT-DNA), often from more than one individual, laboratories have turned to computerized methods to infer which individual profiles may be giving rise to the patterns of STRs and which features in the data are the result of instrumental error or limitations. Both the modifications for generating data from LT-DNA and the "probabilistic genotyping software" have been the subject of admissibility hearings.

Less frequently, other genetic systems to establish the identity of the source of trace DNA evidence have been applied to generate investigative leads. The highly publicized technique of trawling commercial or other privately maintained databases established for genealogical research, to discover individuals who might be related to the source, is discussed in the section titled "Investigative Genetic Genealogy" below. Inferring visible traits and biogeographic ancestry from DNA found at crime scenes is described in the section titled "Forensic DNA Phenotyping" below. These investigative procedures rely on newer methods for analyzing DNA that are just beginning to be evaluated in court.

Throughout these phases, DNA tests also exonerated an increasing number of people who had been convicted of capital and other crimes, posing serious questions about the adequacy of the legal procedures for obtaining postconviction DNA testing and for overturning factually erroneous convictions.[10] The value of

8. PCR (polymerase chain reaction) is described in the section titled "Chromosomes" below.

9. *See* Maryland v. King, 569 U.S. 435 (2013) (upholding compulsory DNA sampling as part of the booking process for custodial arrests); section titled "Law-enforcement databases and databanks" below. Initially, the law-enforcement databases housed records of restriction-fragment-length polymorphisms arising from variable numbers of tandem repeats (RFLP-VNTR profiles). FBI, Frequently Asked Questions on CODIS and NDIS, https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet, last visited Sept. 8, 2024 ("The National DNA Index no longer searches DNA data developed using restriction fragment length polymorphism (RFLP) technology.").

10. *See, e.g.*, Osborne v. Dist. Atty's Office for Third Jud. Dist., 557 U.S. 52 (2009) (narrowly rejecting a convicted offender's claim of a due process right to DNA testing at his expense, enforceable under 42 U.S.C. § 1983, to establish that he is probably innocent of the crime for which he was convicted after a fair trial, when (1) the convicted offender did not seek extensive DNA testing before trial even though it was available, (2) he had other opportunities to prove his innocence after

DNA evidence in solving older crimes also prompted extensions of some statutes of limitations.[11]

In sum, in little more than a decade, forensic DNA typing made the transition from a novel set of methods for identification to a relatively mature and well-studied forensic technology. However, one should not lump all forms of DNA identification together. New techniques and applications continue to emerge. Before admitting DNA evidence, courts normally inquire into the biological principles and knowledge that would justify inferences from new or different technologies or applications. As a result, this guide describes not only the predominant STR technology and earlier methods that were more controversial, but also newer analytical techniques that can be used in human forensic DNA identification.

Relevant Expertise

Human DNA identification can involve testimony about laboratory findings, about the statistical interpretation of those findings, and about the underlying principles of molecular biology and genetics. Consequently, expertise in several fields might be required to establish the admissibility of the evidence or to explain it adequately to the trier of fact. The expert who is qualified to testify about laboratory techniques might not be qualified to testify about molecular biology, to make estimates of population frequencies, or to establish that an estimation procedure is valid.[12]

Trial judges ordinarily are accorded great discretion in deciding when a witness is qualified to testify as an expert, and these decisions depend on the

a final conviction based on substantial evidence against him, (3) he had no new evidence of innocence (only the hope that more extensive DNA testing than that done before the trial would exonerate him), and (4) even a finding that he was not the source of the DNA would not conclusively demonstrate his innocence); Skinner v. Switzer, 562 U.S. 521 (2011); Nat'l Comm'n on the Future of DNA Evidence, Postconviction DNA Testing: Recommendations for Handling Requests (1999); Brandon L. Garrett, *Judging Innocence*, 108 Colum. L. Rev. 55 (2008); Brandon L. Garrett, *Claiming Innocence*, 92 Minn. L. Rev. 1629 (2008).

11. *See, e.g.*, Veronica Valdivieso, Note, *DNA Warrants: A Panacea for Old, Cold Rape Cases?* 90 Geo. L.J. 1009 (2002).

12. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, section titled "Relevant Expertise," in this manual. Nonetheless, if previous cases establish that the testing and estimation procedures are legally acceptable, and if the computations are essentially mechanical, then highly specialized statistical expertise might not be essential. Reasonable estimates of DNA characteristics in major population groups can be obtained from standard references, and many quantitatively literate experts could use the appropriate formulae to compute the relevant profile frequencies or probabilities. NRC II, *supra* note 7, at 170. Limitations in the knowledge of a technician who applies a generally accepted statistical procedure can be explored on cross-examination. *E.g.*, Roberson v. State, 16 S.W.3d 156, 168 (Tex. Crim. App. 2000).

background of each witness. Courts have noted the lack of familiarity of academic experts—who have done respected work in other fields—with the scientific literature on forensic DNA typing and on the extent to which their research or teaching lies in other areas.[13] Although such concerns may affect the persuasiveness of particular testimony, they rarely result in exclusion on the grounds that the witness simply is not qualified as an expert on a particular topic.[14]

Because forensic DNA analysis generally draws on methods and technology developed for research and applications in biological and medical science, the scientific literature on the underpinnings of most forms of DNA evidence, both established and emerging, tends to be extensive. By studying the scientific publications, or perhaps by appointing a special master or expert adviser to assimilate this material,[15] a court can ascertain where a party's expert falls within the spectrum of scientific opinion. Furthermore, an expert appointed by the court under Federal Rule of Evidence 706 could testify about the scientific literature generally or even about the strengths or weaknesses of the particular arguments advanced by the parties.[16]

Given the diversity of forensic questions to which DNA testing might be applied, it is not feasible to list the specific scientific expertise appropriate to all applications. Some applications span many fields. Consider the range of knowledge required to assess the value of DNA analyses of a novel application such as

13. *E.g.*, State v. Copeland, 922 P.2d 1304, 1318 n.5 (Wash. 1996) (noting that defendant's statistical expert "was also unfamiliar with publications in the area," including studies by "a leading expert in the field" whom he thought was "a 'guy in a lab somewhere'").

14. *E.g.*, United States v. Pritchard, 993 F. Supp. 2d 1203, 1208–09 (C.D. Cal. 2014) (laboratory analyst "easily meets the standard for qualification" regarding "statistical analysis testimony" because of her academic education and laboratory and courtroom experience, as supplemented by "special training in statistics" at three or four short courses between 8 and 24 hours in length). *But see* Allen v. State, 62 So. 3d 1199 (Fla. Dist. Ct. App. 2011) (remanding for "a limited evidentiary hearing on the qualifications of the state's expert to testify as to the statistical significance of the DNA profile matches" where witness had taken statistics in college and received on-site training, but the state did not meet its burden of showing the analyst's proficiency and general acceptance of the statistical methodology); Gibson v. State, 915 So. 2d 199 (Fla. Dist. Ct. App. 2005) (although the analyst had taken courses in statistics and had testified that she was following the procedure in the 1996 NRC report, the court of appeals "remanded for a limited evidentiary hearing to determine whether the expert had sufficient knowledge of the authoritative sources to present the statistical evidence"). More case law is collected in George L. Blum, Annotation, *Qualification as Expert to Testify as to Findings or Results of Scientific Test Concerning DNA Matching*, 38 A.L.R. 6th 439 (2008).

15. *See, e.g.,* Gen. Elec. Co. v. Joiner, 522 U.S. 136, 149–50 (1997) (Breyer, J., concurring) (discussing the use of "special masters and specially trained law clerks" in science-related cases); Ass'n of Mex.-Am. Educators v. State of Cal., 231 F.3d 572, 590 (9th Cir. 2000) ("[i]n those rare cases in which outside technical expertise would be helpful to a district court, the court may appoint a technical advisor"); United States v. Lewis, 442 F. Supp. 3d 1122 (D. Minn. 2020) (relying on the report of a special master on the validation and performance of "probabilistic genotyping" software used for the analysis of complex DNA mixtures).

16. *See generally* Kaye et al., *supra* note 5, § 12.2; Faigman et al., *supra* note 5, §§ 1:37 to 1:39.

whole genome sequencing (see section titled "Sequencing and SNPs" below) to establish that a semen sample in a rape case originated from one identical twin rather than the other (see section titled "Twins" below). Expertise in molecular genetics, embryology, biotechnology, bioinformatics, and statistics might be necessary to evaluate the premises and execution of the analysis. Generally, when samples come from crime scenes, the expertise and experience of forensic scientists can be crucial. Just as highly focused specialists may be unaware of aspects of an application outside their field of expertise, so too scientists who have not previously dealt with forensic samples can be unaware of case-specific factors that can confound the interpretation of test results.

Variation in Human DNA and Its Detection

What Are DNA, Chromosomes, Genes, Proteins, and RNAs?

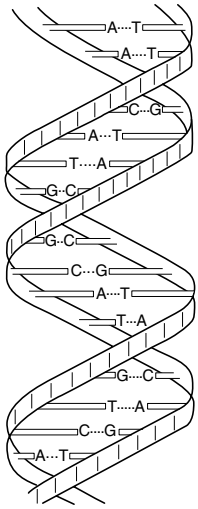
The DNA Molecule

DNA, RNA, and protein molecules are all polymers—large molecules composed of a chain of subunits. The subunits of DNA include four chemical structures known as nucleotide bases. Their names—adenine, thymine, guanine, and cytosine—usually are abbreviated as A, T, G, and C, respectively. The physical structure of DNA is often described as a double helix because the molecule has two spiraling strands connected to each other by weak bonds between the nucleotide bases. As shown in Figure 1, A pairs only with T, and G pairs only with C. Thus, the order of the single bases on either strand reveals the order of the pairs from one end of the molecule to the other, and the DNA molecule represents a long sequence of As, Ts, Gs, and Cs.

Chromosomes

Most human DNA is tightly packed into structures known as chromosomes, which come in different sizes and are located in the nuclei of cells. The chromosomes are numbered, in descending order of size, 1 through 22, with the remaining chromosome being an X or a much smaller Y. Because one of each of these 23 chromosomes is inherited from each parent, the cell's nucleus normally houses 46 chromosomes in all (see the section titled "Sexual Reproduction and the Genome" below). If the bases are like letters, then each chromosome is like a book written in this four-letter alphabet, and the nucleus is like a bookshelf in the

Figure 1. Sketch of a small part of a double-stranded DNA molecule. Nucleotide bases are held together by weak bonds. A pairs with T; C pairs with G.



interior of the cell. All the cells in one individual contain identical copies of the same collection of books.[17] The sequence of the As, Ts, Gs, and Cs that constitutes the "text" in these chromosomes is referred to as the individual's nuclear genome.

All told, the genome has more than three billion "letters" (As, Ts, Gs, and Cs). If these letters were printed in books, the resulting pile would be as high as the Washington Monument. About 99.9% of the genome is identical between any two individuals. This similarity is not really surprising—it accounts for the common features that make modern humans an identifiable species (and for features that we share with many other species as well). The remaining 0.1% is particular to an individual. This variation makes each person (other than identical twins) genetically unique. This small percentage may not sound like a lot, but it adds up to more than three million sites for variation among individuals.

Genes and Gene Products

What are genes, proteins, and RNAs?

Variations in the sequence of base pairs within human populations occur both within the genes and in the regions between the genes. A gene can be defined as

17. Occasional mutations occur, usually as a result of mistakes in the duplication of chromosomes during cell division. *See infra* note 25.

a segment of DNA, usually from 1,000 to 10,000 base pairs long, that "encodes" a protein or an RNA molecule. For example, a large gene named GC (for group-specific component) encodes a protein that binds to vitamin D and is found in blood plasma. In many people, the following tiny sequence appears within that gene:

G C A A A A T T G C C T G A T G C C A C A C C C A A G G A A C T G G C A.

Unlike double-stranded DNA, RNA is a single helical strand that has a different nucleotide (U, which is short for uracil) in place of the T in the DNA helix. As indicated below, RNA plays an integral part in gene expression.

Proteins are composed of units called amino acids. Twenty different amino acids exist in proteins, and hundreds to thousands of them are attached to each other in long chains with more complex shapes than the single helix of RNA or the double helix of DNA. Proteins perform all sorts of functions in the body and thus produce observable characteristics. For example, the group-specific component protein referred to above (and designated Gc) grabs onto vitamin D and circulates it in the body. It also plays a role in the immune system.

The order of the building blocks of this protein can be slightly different in different individuals. Forensic scientists used these differences before DNA testing was available to exclude or include suspects or defendants as possible sources of blood or semen stains. The cell produces specific proteins and RNAs that correspond to the order of the bases (the "letters") in the coding part (exons) of a gene. The sequence in which the protein's amino acids are arranged corresponds to the sequence of base pairs within a gene. A sequence of three base pairs (a codon) specifies a particular 1 of the 20 possible amino acids in the protein. The mapping of various sequences of three nucleotide bases to a particular amino acid is the genetic code. About 1.5% of the human genome codes for the amino-acid sequences.

Human genes also contain noncoding sequences that regulate the cell type in which a protein will be synthesized and how much protein will be produced. Many genes contain interspersed noncoding, nonregulatory sequences that no longer participate in protein synthesis. These sequences, called introns, constitute about 23% of the base pairs within human genes. In terms of the metaphor of DNA as text, the gene is like an important paragraph in the book, often with some gibberish in it.[18]

18. The idea of a gene as a block of DNA (some of which is coding, some of which is regulatory, and some of which is functionless) is an oversimplification (*see, e.g.,* Petter Portin & Adam Wilkins, *The Evolving Definition of the Term "Gene,"* 205 Genetics 1353 (2017)), but it is useful enough here.

How do cells manufacture proteins and RNAs?

Protein-coding genes express their sequence-specific information by an intricate process.[19] Through a series of biochemical reactions, the base pairs of the exons of the gene are transcribed into an RNA molecule. The noncoding parts (the introns) are not represented in this messenger RNA (mRNA) transcript. The transcribed mRNA makes its way to microscopic molecular factories, where it is translated into a protein.[20] By combining the amino acids in different orders, a virtually unlimited variety of proteins can be constructed. Other genes contain DNA sequences that are transcribed into RNAs that perform other functions. Their final products are nonprotein-coding RNAs (ncRNAs).[21]

Alleles and Loci of Genes

The GC gene mentioned above always is located at the same position, on chromosome 4. But the short sequence listed as occurring within that gene is not the same for every individual. A locus where almost all humans have the same DNA sequence is called monomorphic ("of one form"). A locus where the DNA sequence varies among significant numbers of individuals—more than 1% or so of the population possesses the variant—is called polymorphic ("of many forms"). The alternative forms are called alleles. The GC locus, for example, is

19. The description here of how genes express proteins and RNAs is adapted, in part, from a more complete explanation in the Brief of Genetics, Genomics and Forensic Science Researchers as Amici Curiae in Support of Neither Party in *Maryland v. King*, 569 U.S. 435 (2013) (reprinted in Henry T. Greely & David H. Kaye, *A Brief of Genetics, Genomics and Forensic Science Researchers in* Maryland v. King, 53 Jurimetrics J. 43 (2013)) [hereinafter Amici Brief].

20. Transfer RNA brings the building blocks of proteins (the amino acids harvested from food) to these structures. There, they are joined in the order dictated by the mRNA transcript.

21. The ncRNAs include RNAs that do the work of translation and RNAs involved in regulating expression of dozens or even hundreds of protein-coding genes. Furthermore, much shorter RNAs regulate transcription and translation. Thus, it is widely recognized that the genome is abuzz with transcription-to-RNA activity and other events that interact in the expression of the protein-coding DNA. The discoveries of these processes generated speculation in the press and some judicial opinions that *all* DNA sequences significantly affect development and health. The *King* amici maintained that

> This is not true—even for sequences that are transcribed. Not every biochemical event along the DNA has a measurable impact on health. First, some short ncRNA transcripts are just "noise." They are degraded quickly. Second, intronic parts of the pre-mRNA transcripts normally are removed. Third, even if one transcript could regulate expression, other transcripts may do the same job. Fourth, even if the stable transcript does affect some trait, the effect may be so minor in the context of other genetic and environmental influences as to be of no meaningful predictive or diagnostic value. Finally, even if a stable transcript has a dramatic effect on a trait, that trait may be unrelated to disease status. Consequently, regarding every bit of the genome as if it were a medical record is unwarranted.

Amici Brief, *supra* note 19, at 11.

polymorphic. The sequence has three common alleles that result from substitutions in a base at a given point. Where an A appears in one allele, there is a C in another. The third allele has the A, but at another point a G is swapped for a T. These changes are called single nucleotide polymorphisms (SNPs, pronounced "snips").

If a gene is like a paragraph in a book, a SNP is a change in a letter somewhere within that paragraph (a substitution, a deletion, or an insertion), and the two versions of the gene that result from this slight change are the alleles. An individual who inherits the same allele from both parents is called a homozygote. An individual with distinct alleles is a heterozygote.

DNA sequences used for forensic analysis usually are not inside the coding parts of genes. They lie in the vast regions between genes (about 75% of the genome is extragenic) or in the introns. These extra- and intragenic regions of DNA have been found to contain considerable sequence variation, which makes them particularly useful in distinguishing individuals. Although the terms "locus," "allele," "homozygous," and "heterozygous" were developed to describe genes, the nomenclature has been carried over to describe all DNA variation—coding and noncoding alike. Both types are inherited from mother and father in the same fashion, as discussed in the next subsection.

Sexual Reproduction and the Genome

The process that gives rise to the diversity of alleles starts with the production of special sex cells—sperm cells in males and egg cells in females. All the nucleated cells in the body other than sperm and egg cells contain two versions of each of the 23 chromosomes—two copies of chromosome 1, two copies of chromosome 2, and so on, for a total of 46 chromosomes. The 22 numbered chromosomes are called autosomes. Cells in females contain two X chromosomes, and cells in males contain one X and one Y chromosome.[22] An egg cell, however, contains only 23 chromosomes—one chromosome 1, one chromosome 2, . . . , one chromosome 22, and one X chromosome—each selected at random from the woman's full complement of 23 chromosome pairs. Thus, each egg carries half the genetic information present in the mother's 23 chromosome pairs—a "haploid" genome. Because the assortment of the chromosomes is random, each egg carries a different complement of genetic information. Further randomness comes from a process known as crossing over, in which segments of each of the mother's two paired chromosomes exchange places in producing the single chromosome from the pair that ends up in

22. A small fraction of people are born with extra chromosomes or a missing one. The most common such condition is Down syndrome (an extra chromosome 21), which occurs in about 1 in every 700 babies born. Nat'l Center on Birth Defects & Developmental Disabilities, Centers for Disease Control and Prevention, Data and Statistics on Down Syndrome (Dec. 16, 2022), https://perma.cc/VB5W-XM5D.

the egg. The same situation exists with sperm cells. Each sperm cell contains a single copy of each of the 23 chromosomes (with crossing over) selected at random from a man's 23 pairs.[23] Fertilization of an egg by a sperm therefore restores the full number of 46 chromosomes, with the 46 chromosomes in the fertilized egg—a "diploid" genome—that is a new combination of those in the mother and father. This recombination of genetic information in sexual reproduction is the main source of human genetic diversity.

During pregnancy, the fertilized cell divides to form two cells, each of which has an identical copy of the 46 chromosomes. The two then divide to form four, the four form eight, and so on. As gestation proceeds, various cells specialize (differentiate) to form different tissues and organs. Although cell differentiation yields many different kinds of cells, the process of cell division usually results in a line of cells having the same genomic complement as the cell that divided. Thus, aside from occasional mutations occurring after fertilization,[24] each of the approximately 100 trillion cells in the adult human body has the same DNA as was present in the original 23 pairs of chromosomes from the fertilized egg, one member of each pair having come from the mother and one from the father.[25]

23. The man's two sex chromosomes (an X and a Y) are so different that they do not undergo crossing over during the formation of a sperm cell (except in small "pseudo-autosomal regions"—regions at one or both ends of the pair).

24. Mutations in a parent's sex cell (an egg or sperm) that occur before conception will be incorporated into the genome of the progeny by this repeated copying process. Such prefertilization mutations could complicate parentage or other relationship testing, but they do not matter when testing to see whether two samples originated from the same progeny—for example, a suspect in a criminal case.

25. Mutations can occur during embryonic development or later. The most important source of these changes is a mistake in the DNA copying process during cell division. These are low probability events, but given the immense number of cell divisions that occur during the life of an organism, some are to be expected. Thus, every human being may well be a genetic mosaic to some degree. The mutation rate seems to vary across sites within the genome and between different cell types. Some mutations result in cancers—cells whose proliferation is not inhibited by normal mechanisms. Cristian Tomasetti et al., *Stem Cell Divisions, Somatic Mutations, Cancer Etiology, and Cancer Prevention*, 355 Science 1330 (2017), https://doi.org/10.1126/science.aaf9011. New cell lineages arising after birth (cancerous or otherwise) do not normally confuse forensic DNA comparisons because the vast majority of them would not occur at the loci used in identity testing (*see* section titled "What Are DNA Polymorphisms and How Are They Detected?" below), and even if they did, the fraction of the mutated cells in most forensic DNA samples would be so small that only the unmutated alleles would be detected. It is also possible that a mutation would occur "so early in embryonic development that DNA in eggs or sperm might differ from that in blood [or saliva] from the same person." NRC II, *supra* note 7, at 64 n.7. The NRC report characterizes this mosaicism as "a remote possibility." *Id*. Whether or not this is accurate, as with mutations occurring later in life, the embryonic mutations resulting in the divergence between tissues would have to change the specific loci used in forensic identity testing, which is additionally improbable.

What Are DNA Polymorphisms and How Are They Detected?

By determining which alleles are present at strategically chosen locations on a chromosome (loci), the forensic scientist ascertains the DNA profile, or genotype, of an individual (at those loci). Although the differences among the alleles arise from alternations in the order of nucleotide base pairs (the A, T, G, C letters), DNA typing for ascertaining identity does not require "reading" the full genome. Here we outline the major types of polymorphisms that have been used in identity testing and the methods for detecting them.

VNTRs and RFLP Testing

The first polymorphisms to find widespread use in identity testing result from the presence of a variable number of tandem repeats (VNTRs) at a locus. Being the subject of most of the court opinions on the admissibility of DNA evidence in the late 1980s and early 1990s, they paved the way for the methods now in use. These VNTRs are one type of repetitive DNA. They are like a musical score in which the same melody is played over and over without interruption. The core unit (the melody) is a particular short DNA sequence, and the number of times it is repeated varies from one person to another. The first VNTRs to be used in genetic and forensic testing had core repeat sequences of 7–35 or more base pairs.

In forensic VNTR testing, enzymes (known as restriction enzymes) were used to cut the DNA molecule both before and after the VNTR sequence. A small number of repeats in the VNTR region gives rise to a small restriction fragment, and a large number of repeats yields a large fragment. A substantial quantity of DNA is required to give a detectable number of VNTR fragments with this procedure. After the restriction fragments are sorted by size with a process known as gel electrophoresis,[26] the fragments with particular VNTRs are detected by applying a "probe" that binds when it encounters the repeated core sequence. A probe is a short, synthesized fragment of single-stranded DNA with base pairs arranged to complement the core sequence. For example, if the core's sequence is AGTTCGCGTGACTAT, the probe's sequence would be TCAAGCGCACGATA. Because A bonds to T and G to C (and vice versa; see Figure 1), when the probe comes in contact with the restriction fragment, it sticks to it. A radioactive

26. Electrophoresis separates DNA, RNA, or protein molecules according to their size. An electric current drags the molecules through a gel or other matrix. Smaller molecules move through the matrix more quickly than larger molecules. Molecules of known sizes ("standards") are run through the gel at the same time as the molecules from the sample. Comparing their positions at the end of the run to those of the separated molecules indicates the sizes of the latter.

or fluorescent molecule attached to the probe provides a way to mark the VNTR fragment. Many court opinions refer to this process as RFLP testing.[27]

PCR Amplification

Most modern forensic-science methods of DNA analysis take advantage of a chemical reaction called PCR (for polymerase chain reaction). The PCR process enables laboratories to make many copies of small DNA fragments.[28] Other procedures then can be applied to analyze the vastly amplified quantity of DNA.

PCR can be applied to the double-stranded DNA segments extracted and purified from a forensic sample as follows: First, the purified DNA is separated into two strands by heating it to near the boiling point of water.[29] Second, the single strands are cooled, and primers—synthesized fragments of single-stranded DNA, usually between 15 and 30 nucleotides long—attach themselves to the points at which the copying will start and stop.[30] Finally, the soup containing the annealed DNA strands, an enzyme called DNA polymerase, and lots of the four nucleotide building blocks are warmed. On a DNA strand, the polymerase inserts the complementary bases one at a time, building a new strand bound to the original template and thus replicating part of the DNA strand that was separated from its partner in the first step. The same replication occurs with the separated partner as the template.[31] The result is two identical double-stranded DNA segments, one made from each strand of the original DNA. The three-step cycle is repeated, usually 20 to 35 times in automated machines known as thermal cyclers. Ideally, starting with a single double-stranded molecule, the first cycle results in two double-stranded DNA segments; the second cycle produces four; the third, eight; and so on, until there are millions of copies of the DNA sequences of interest.[32]

Care must be taken to achieve the appropriate chemical conditions. Furthermore, because even small amounts of stray DNA could be amplified along with the DNA of interest, quality-assurance steps must be taken to reduce contamination

27. It would be clearer to call it RFLP-VNTR testing, because the fragments being measured contain the VNTRs rather than some simpler polymorphisms that were used in genetic research and disease testing. A more detailed exposition of the steps in RFLP-VNTR profiling (including gel electrophoresis, Southern blotting, and autoradiography) can be found in the first edition of this guide (Reference Manual on Scientific Evidence (1st ed. 1994)) and in many judicial opinions circa 1990.

28. Most VNTRs are too long for PCR to work efficiently.

29. This "denaturing" takes about a minute. Chemical and other physical methods for denaturing also can be used.

30. By judiciously constructing the primers to bracket the sequences of interest, the amplified DNA will consist almost entirely of the in-between sequences. The rest of the genome will not be copied. Annealing the primers takes about 45 seconds.

31. This extension step for both templates takes about two minutes.

32. In practice, there is some inefficiency in the doubling process, but the yield from a 30-cycle amplification is generally about 1 million to 10 million copies. NRC II, *supra* note 7, at 69–70.

of the sample.[33] A laboratory should be able to demonstrate that it can amplify targeted sequences faithfully with the equipment and reagents that it uses and that it has taken suitable precautions to minimize or detect contamination from extraneous DNA.[34] With small samples, it is possible that some alleles will be amplified and others missed (preferential amplification, discussed below in the section titled "Did the Sample Contain Enough DNA?"). Mutations within a population giving rise to variants in the region of a primer can prevent the amplification of the allele downstream of the primer, giving rise to null alleles.[35]

STRs and Capillary Electrophoresis

How STRs differ from VNTRs

Although RFLP-VNTR profiling is highly discriminating,[36] it has several drawbacks. Not only does it require a substantial sample of DNA-bearing material, but it is time-consuming and does not measure the fragment lengths to the nearest number of repeats.[37] Consequently, forensic scientists moved from VNTRs to another form of repetitive DNA known as short tandem repeats

33. In some instances, interpretation of results may be valuable notwithstanding known contaminating DNA. For example, it may be possible to exclude an individual as a viable source. *See* Human Forensic Biology Subcomm., Organization of Scientific Area Committees for Forensic Science, Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events, June 1, 2021 (OSAC 2020-S-0004).

34. Forensic-science organizations have proposed guidelines. *See, e.g.,* Eur. Network Forensic Sci. Inst., Guideline for DNA Contamination Minimization in DNA Laboratories, May 10, 2023, https://perma.cc/EQB4-55UB. As have researchers and organizations concerned with PCR-based testing in clinical and research laboratories. *E.g.*, World Health Organization, Dos and Don'ts for Molecular Testing (Jan. 31, 2018), https://perma.cc/LK5B-NBC3.

35. In the context of STR profiling (discussed in the next section), a "null allele is any allele at a microsatellite [STR] locus that consistently fails to amplify to detected levels via the polymerase chain reaction." Elizabeth E. Dakin & John C. Avise, *Microsatellite Null Alleles in Parentage Analysis*, 93 Heredity 504, 504 (2004), https://doi.org/10.1038/sj.hdy.6800545. Such a null allele will not lead to a false exclusion if the two DNA samples from the same individual are amplified with the same primer system, but it could lead to an exclusion at one locus when searching a database of STR profiles if the database profile was determined with a different PCR kit than the one used for the crime-scene DNA.

36. Alleles at VNTR loci generally are too long to be measured precisely by electrophoretic methods—alleles differing in size by only a few repeat units may not be distinguished. Although this makes for complications in deciding whether two length measurements that are close together result from the same allele, these loci are quite powerful for the genetic differentiation of individuals, because they tend to have many alleles that occur relatively rarely in the population. At a locus with only 20 such alleles (and most loci typically have many more), there are 210 possible genotypes. With 5 such loci, the number of possible genotypes is 210^5, which is more than 400 billion.

37. The measurement error inherent in the form of electrophoresis used is not a fundamental obstacle, but it complicates the determination of which profiles match and how often other profiles in the population would be declared to match. For a case reversing a conviction as a result of an

(STRs). STRs have very short core repeats, two to seven base pairs in length, and they typically extend for only some 50 to 350 base pairs.[38] Like the larger VNTRs, which extend for thousands of base pairs, STR sequences do not code for proteins or RNAs, and the ones routinely used in identity testing are four-nucleotide repeats thought to have little or no clinical value in ascertaining an individual's disease status, propensity, or behavioral characteristics.[39] The STR loci used in forensic identification have names such as TPOX[40] and D16S539. Figure 2 illustrates the nature of allelic variation at the latter locus, on chromosome 16.[41] There are other forensic STR loci with more complicated internal patterns.[42] Although there are fewer alleles per locus for STRs than for VNTRs, many more STRs are analyzed simultaneously (see section titled "Capillary electrophoresis and fluorescence" below).

Figure 2. Three alleles of the D16S539 STR.

Nine-repeat allele:
GATAGATAGATAGATAGATAGATAGATAGATAGATA
Ten-repeat allele:
GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA
Eleven-repeat allele:
GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA

Note: The core sequence is GATA. The first allele listed has nine tandem repeats, the second has ten, and the third has eleven. The locus has other alleles (different numbers of repeats), shown in Figure 4.

expert's confusion on this score, see *People v. Venegas*, 954 P.2d 525 (Cal. 1998). More suitable procedures for match windows and probabilities are described in NRC II, *supra* note 7.

38. The numbers, and the distinction between "minisatellites" (VNTRs) and "microsatellites" (STRs), are not precise, but the mechanisms that give rise to the shorter tandem repeats differ from those that produce the longer ones. *See* Jocelyn E. Krebs et al., Lewin's Genes XII (2018).

39. *See* Amici Brief, *supra* note 19; Sara H. Katsanis & Jennifer K. Wagner, *Characterization of the Standard and Recommended CODIS Markers*, 58 J. Forensic Sci. S169 (2013), https://doi.org/10.1111/j.1556-4029.2012.02253.x; David H. Kaye, *Please, Let's Bury the Junk: The CODIS Loci and the Revelation of Private Information*, 102 Nw. U. L. Rev. Colloquy 70 (2007), *available at* https://perma.cc/8QA3-Y3ES. *But see* Mayra M. Bañuelos et al., *Associations Between Forensic Loci and Expression Levels of Neighboring Genes May Compromise Medical Privacy*, 119 Proc. Nat'l Acad. Sci. e2121024119 (2022), https://doi.org/10.1073/pnas.2121024119; Nicole Wyner et al., *Forensic Autosomal Short Tandem Repeats and Their Potential Association with Phenotype,* 11 Frontiers Genetics 884 (2020), https://doi.org/10.3389/fgene.2020.00884 (concluding that no "forensic STRs . . . were found to be independently causative or predictive of disease," but proposing that "there remains a strong chance that this inference may change in the near future").

40. This STR is found in an intron of the thyroid peroxidase gene.

41. Names of the STRs not found in genes are in the form D#S#. The first number indicates the chromosome; the second, a location on the chromosome.

42. For examples, see Peter Gill et al., *Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles* 2–3 (2020).

Medical and human geneticists were interested in VNTRs and STRs as markers in family studies to locate the genes that are associated with inherited diseases. Papers on the potential for identity testing appeared in the early 1990s. Developmental research to pick suitable loci moved into high gear in England and other parts of Europe. Britain's Forensic Science Service applied a four-locus testing system in 1994. Then it introduced the second-generation multiplex (SGM)—for simultaneously typing six loci in 1996. These soon would be used to build the U.K.'s National DNA Database. The database system allows a computer to check the STR types of millions of known or suspected criminals against thousands of crime-scene samples. A six-locus STR profile can be represented as a string of 12 digits; each digit indicates the number of repeat units in the alleles at each locus. These discrete, numerical DNA profiles are far easier to compare mechanically than the complex patterns of fingerprints. In the United States, the FBI settled on 13 "core loci" to use in the U.S. national DNA database system.[43] This number is capable of distinguishing among almost everyone in the population.[44] These are often called the CODIS core loci, and an additional seven STR loci were added to the system in 2017.[45] The remainder of this section describes how laboratories typically determine which STR alleles are present after DNA from the sample is isolated.

Capillary electrophoresis and fluorescence

Determining which STR alleles are present involves ascertaining the size of the fragments that have been amplified at each STR locus. Separation according to fragment length is done in automated "genetic analyzer" machinery—a byproduct of the technology developed for the Human Genome Project that first sequenced most of the entire genome. In these machines, a long, narrow tube is filled with an entangled polymer or comparable sieving medium, and an electric

43. The federally managed Combined DNA Index System (CODIS) and related issues are discussed in the section titled "Offender and Suspect Database Searches" below; *see also* John M. Butler, Advanced Topics in Forensic DNA Typing: Methodology 21370 (2012).

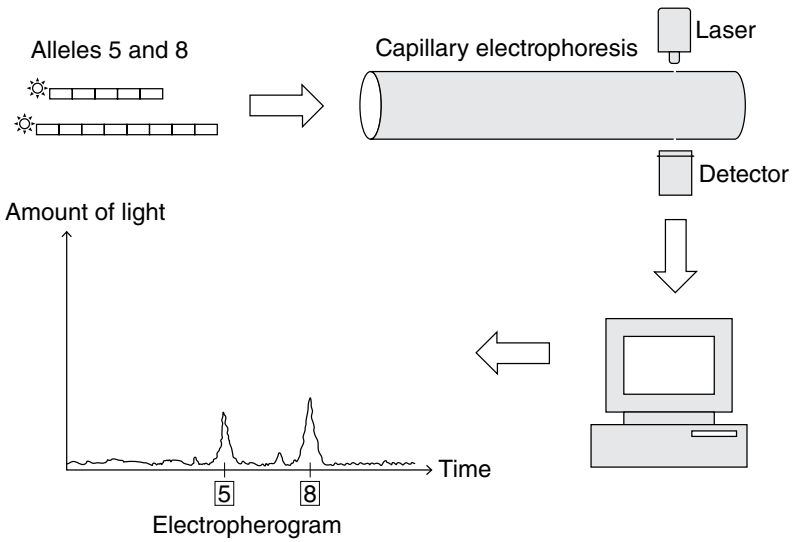
44. Usually, there are between 7 and 15 STR alleles per locus. Thirteen loci that have 10 STR alleles each can give rise to 55^{13}, or 42 billion trillion, possible genotypes—far, far greater than the U.S. population of 330 million or so. The enormous number of possible 13-locus profiles suggests that it is improbable—but not necessarily impossible—for two unrelated individuals in the country to have identical ones. Even first-degree relatives are unlikely to have the same alleles at all 13 loci. *See* section titled "Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing" below. But identical twins are expected to share the same STR profile. *See* section titled "Twins" below.

45. Douglas R. Hares, *Selection and Implementation of Expanded CODIS Core Loci in the United States*, 17 Forensic Sci. Int'l: Genetics 33 (2015), https://doi.org/10.1016/j.fsigen.2015.03.006. CODIS stands for "combined DNA index system" (discussed in section titled "Offender and Suspect Database Searches" below).

field is applied to pull DNA fragments placed at one end of the tube through the medium. As with gel electrophoresis of RFLPs (see section titled "VNTRs and RFLP Testing" above), shorter fragments slip through the medium more quickly than larger, bulkier ones.

Detecting the fragments as they pass through the capillary tubes is made possible by attaching a fluorescent dye molecule to the PCR primer. As the amplified fragments move through the medium, a laser beam is sent through a small glass window in the tube, causing the dye to glow at a characteristic wavelength. The intensity of the fluorescence is recorded by a kind of electronic camera and transformed into a graph (an electropherogram), which shows a peak as the fragments with the fluorescing dye flash by. Copies of a shorter allele will pass by the window and fluoresce first; copies of a longer fragment will come by later, giving rise to another peak on the graph. Figure 3 is a sketch of how the alleles with five and eight repeats of the GATA sequence at the D16S539 STR locus might appear in an electropherogram.

Figure 3. Sketch of an electropherogram for two D16S539 alleles.



Note: One allele has five repeats of the sequence GATA; the other has eight. Each GATA repeat is depicted as a small rectangle. Although only one copy of each allele (with a fluorescent molecule, or "tag" attached) is shown here, PCR generates a great many copies from the DNA sample with these alleles at the D16S539 locus. These copies are drawn through the capillary tube, and the tags glow as the STR fragments move through the laser beam. An electronic camera measures the colored light from the tags. Finally, a computer processes the signal from the camera to produce the electropherogram.
Source: David H. Kaye, The Double Helix and the Law of Evidence 189, fig. 9.1 (2010).

Modern genetic analyzers produce electropherograms for many loci at once.[46] This "multiplexing" is accomplished by using dyes that fluoresce at distinct colors to label the alleles from different groups of loci. A separate set of fragments of known sizes that comigrate through the capillary function as a kind of ruler (an internal-lane size standard) to determine the lengths of the allelic fragments. Software processes the raw data to generate an electropherogram of the separate allele peaks of each color. By comparing the positions of the allele peaks to the size standard, the program determines the number of repeats in each allele. The plotted heights of the peaks (measured in relative fluorescent units, or RFUs) are proportional to the amount of the PCR product.

Complications can arise, especially with minute samples and mixtures. During PCR, the DNA strand being copied and the new strand can slip, causing "stutter" peaks that typically are one repeat away from the originating STR alleles. Fragments of DNA can fall into PCR tubes or be carried by dust particles, causing "drop in" alleles; color dyes do not fluoresce solely at the desired color and can give a signal in the adjacent color channel—a spectral artifact known as pull-up; a variety of factors can lead to unequal peaks for the pair of alleles at a locus (peak imbalance) and to missing peaks (allele dropout or even whole locus dropout).[47]

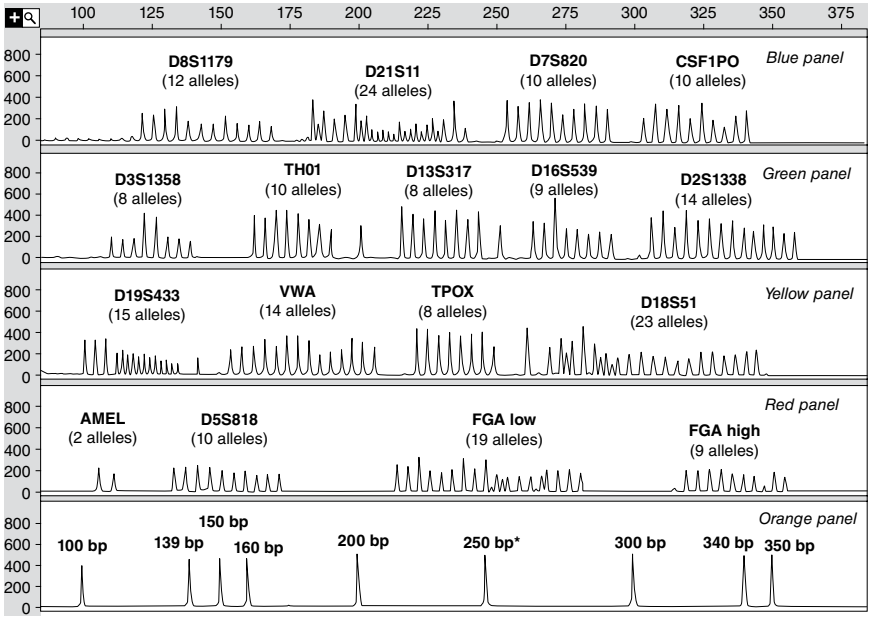
Figure 4 is an electropherogram of all 203 major alleles at 15 STR loci typed in a single multiplex PCR reaction. In addition, it shows the two alleles of the gene used to determine the biological sex of the contributor of a DNA sample.[48] An electropherogram from an individual's DNA would have only one or two peaks at each of these 15 STR loci (depending on whether the person is homozygous or heterozygous). These "allelic ladders" aid in deciding which allele a peak from an unknown sample represents.

46. Kathryn Oostdik et al., *Developmental Validation of the PowerPlex® Fusion System for Analysis of Casework and Reference Samples: A 24-locus Multiplex for New Database Standards*, 12 Forensic Sci. Int'l Genetics 69 (2014), https://doi.org/10.1016/j.fsigen.2014.04.013; Suhua Zhang et al., *Development and Validation of a New STR 25-plex Typing System*, 17 Forensic Sci. Int'l: Genetics 61 (2015), https://doi.org/10.1016/j.fsigen.2015.03.008.

47. *See also* sections titled "Did the Sample Contain Enough DNA?" and "Mixtures" below.

48. The amelogenin gene, which is found on the X and the Y chromosome, codes for a protein that is a major component of the tooth enamel matrix. The copy on the Y chromosome is 112 base pairs (bp) long. The copy on the X chromosome has a string of six base pairs deleted, making it slightly shorter (106 bp). A female (XX) will have one peak at 112 bp. A male (XY) will have two peaks (at 106 and 112 bp). In some populations, however, mutations that interfere with the detection of the Y chromosome by this method have been observed. These could lead to male DNA being misidentified as female on the basis of this locus alone. *See* Vijendra Kumar Kashyap et al., *Deletions in the Y-derived Amelogenin Gene Fragment in the Indian Population*, 7 BMC Med. Genetics 37 (2006), https://doi.org/10.1186/1471-2350-7-37.

Figure 4. Alleles of 15 STR loci and the amelogenin sex-typing test from the AmpFlSTR Identifiler kit.

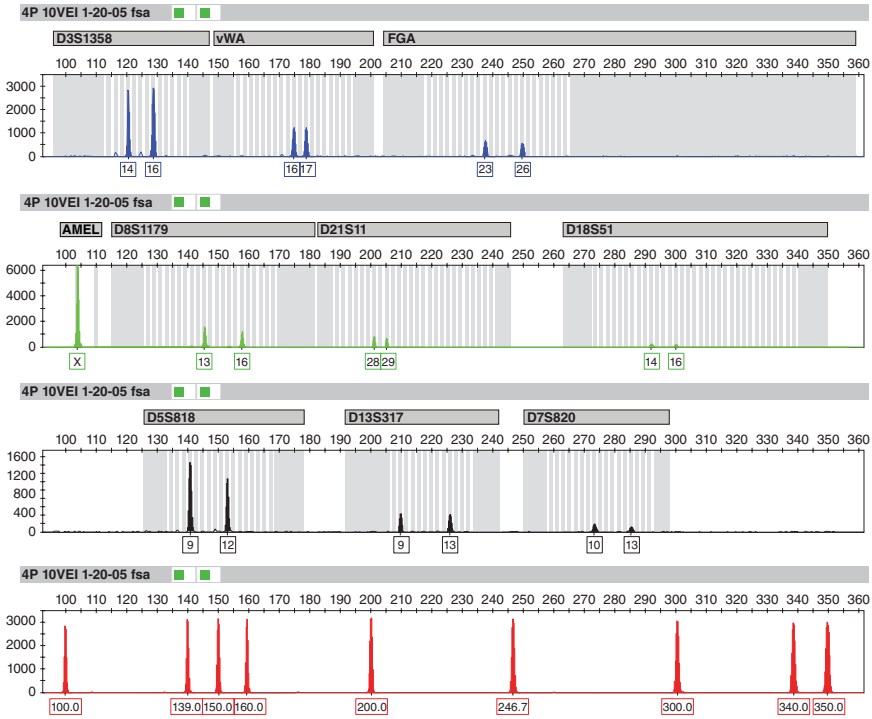


Note: The bottom panel is a sizing standard—a set of peaks from DNA sequences of known lengths (in base pairs). The numbers in the vertical axis in each panel are relative fluorescence units (RFUs) that indicate the amount of light emitted after the laser beam strikes the fluorescent tag on an STR fragment. Applied Biosystems makes the kit that produced these allelic ladders.
Source: John M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers 128 (2d ed. 2005). Copyright Elsevier 2005, with the permission of Elsevier Academic Press. John Butler supplied the illustration.

Figure 5 is an electropherogram from the vaginal epithelial cells of the body of a girl who had been sexually assaulted and killed in California.[49] It was produced for the retrial in 2008 of the defendant, who was linked to the victim by VNTR typing at his first trial in 1990.

49. People v. Pizarro, 12 Cal. Rptr. 2d 436 (Ct. App. 1992), *after remand*, 3 Cal. Rptr. 3d 21 (Ct. App. 2003), *review denied* (Oct. 15, 2003).

Figure 5. Electropherogram for nine STR loci of the victim's DNA in *People v. Pizzaro*.



Note: The amelogenin locus and a sizing standard at the bottom also are included. Some STR loci have small peaks, indicating that there was not much PCR product for those loci, likely because of DNA degradation. All of the STR loci have two peaks, as would be expected when the source is heterozygous at those loci.
Source: Steven Myers and Jeanette Wallin, California Department of Justice, provided the image.

Miniaturization for rapid capillary electrophoresis

Miniaturized capillary electrophoresis (CE) devices have been developed for rapid detection of STRs and other genetic analyses. The mini-CE systems consist of microchannels etched on glass, silicon, or plastic wafers ("chips") using technology borrowed from the computer industry. The microchannels are roughly the diameter of a hair. The principles of electrophoretic separation are the same as with conventional CE systems, but with microfluidic technologies, it is possible to integrate DNA extraction and PCR amplification processes with the CE separation in a single device, a so-called lab on a chip instrument that includes a cartridge with miniature pumps, mixers, valves, fluid channels

and reagent storage chambers.[50] Once a sample is added to the cartridge, all the analytical steps are performed without further human contact. These devices combine simplified sample handling with rapid analysis.

"Rapid DNA" devices have been commercialized for point-of-care medical diagnostics[51] and for military and police applications. Validation for police work can be achieved by having police obtain profiles under realistic conditions and then re-running the samples with conventional equipment at established laboratories; or samples from known sources of DNA can be created for testing under experimental conditions.[52] Several different rapid DNA machines for forensic STR profiling have been used by police departments or other agencies across the world in various ways: (1) to acquire a DNA profile during an initial encounter with a suspect; (2) to obtain the profile quickly after arrest as part of the booking process; (3) to extract and analyze DNA from crime-scene and sexual-assault-kit samples; and (4) to identify victims of mass disasters.[53] In the first situation, police may lack probable cause to arrest an individual but may try to conform to the requirements of the Fourth Amendment by securing valid consent to the DNA profiling.[54] Although some local police agencies—and a large prosecutors' office—have pursued this

50. Lab-on-a-chip instruments (often called LOCs) also can use other methods than CE for analyzing DNA, proteins, or other molecules. *See, e.g.,* Prapti Pattanayak et al., *Microfluidic Chips: Recent Advances, Critical Strategies in Design, Applications and Future Perspectives*, 25 Microfluidics & Nanofluidics 99 (2021), https://doi.org/10.1007/s10404-021-02502-2.

51. Curtis D. Chin et al., *Commercialization of Microfluidic Point-of-Care Diagnostic Devices*, 12 Lab on a Chip 2118 (2012), https://doi.org/10.1039/c2lc21204h; P. Yager et al., *Microfluidic Diagnostic Technologies for Global Public Health*, 442 Nature 412 (2006), https://doi.org/10.1038/nature05064. However, those applications would not use human STRs; instead, they would detect the DNA or RNA of pathogens.

52. *E.g.*, Rosemary S. Turingan et al., *Developmental Validation of the ANDE 6C System for Rapid DNA Analysis of Forensic Casework and DVI Samples*, 65 J. Forensic Sci. 1056 (2020), https://doi.org/10.1111/1556-4029.14286. More detailed guidelines for validation can be found in the United Kingdom Forensic Science Regulator's report, *Forensic Science Regulator Guidance: Methods Employing Rapid DNA Devices* § 8 (2021), https://perma.cc/746P-9XW3 (FSR-G-229).

53. The current systems take about 90 minutes to analyze a sample.

54. Some localities go further and keep the STR data in a computer-searchable, local DNA database that they can operate independently of the databases that are part of the system of shared profile data (known as CODIS), which are administered by the FBI pursuant to the DNA Identification Act of 1994, 34 U.S.C. §§ 40702–40703 (later amended in various respects) and state statutes. Whether the *indefinite* retention and trawling of the local database of consent profiles for matches to future crime-scene samples exceeds the scope of the consent in a given case could be a significant question. Moreover, the non-CODIS local databases (most of which were created without rapid DNA instruments) have been criticized on other grounds. *See, e.g.*, Jason Kreag, *Going Local: The Fragmentation of Genetic Surveillance*, 95 B.U. L. Rev. 1491 (2015); N.Y. City Bar Ass'n Crim. Courts Comm., Crim. Just. Operations Comm., and Mass Incarceration Task Force, Curbing Unregulated Local DNA Indexing, Apr. 16, 2021 (report supporting "an Act . . . requiring municipalities to expunge any DNA record stored in a municipal DNA identification index," accessible via https://perma.cc/72FC-PGL7). In Arizona, the state Department of Public Safety chose to establish a non-CODIS database, relying on rapid DNA instruments and DNA database

practice,[55] rapid DNA technology has been more widely adopted as an adjunct to arrests. Pursuant to the Rapid DNA Act of 2017,[56] the FBI issued standards and procedures for the use of rapid DNA instruments in the booking station that allow the resulting DNA profiles to be uploaded to the National DNA Index System.[57] It is working cautiously toward the third approach, of rapid DNA analysis of crime-scene and sexual-assault-victim DNA samples.[58]

Rather than face objections to rapid DNA based on Rule 702, prosecutors might choose to have accredited laboratories redo the analyses that police perform and introduce only the latter evidence. This will be possible for fresh DNA samples from defendants, but not when the crime-scene DNA has been totally consumed during the rapid testing. As with the widespread use of Breathalyzers for blood-alcohol testing,[59] ultimately courts will confront Rule 702 issues of scientific validity and proper use of the equipment. A number of organizations and individuals have expressed reservations about the current state of the technology,[60] and research is ongoing.[61]

management software purchased from vendors, for the benefit of localities that otherwise might have been tempted to establish such systems on their own. Kreag, *supra,* at 1517–19.

55. In Orange County, California, prosecutors with rapid DNA machines in their basement assembled a significant database by "offer[ing] defendants accused of misdemeanors and infractions a deal: give the prosecutor's office your DNA, and the office will offer you leniency in your criminal case." Andrea Roth, *"Spit and Acquit": Prosecutors as Surveillance Entrepreneurs*, 107 Cal. L. Rev. 405, 408 (2019). California Senate Bill 1228 (adopted 2022) curbs such practices.

56. Pub. L. No. 115–50, 131 Stat. 1001 (codified at 34 U.S.C. §§ 12591, 12592, 40702, 40703).

57. *See* FBI, Guide to All Things Rapid DNA, Jan. 27, 2022, https://le.fbi.gov/file-repository/rapid-dna-guide-january-2022.pdf/view. Previously, only qualified state laboratories could submit profiles.

58. *Id.*

59. *See* 1 McCormick on Evidence, *supra* note 1, § 205.1, at 1309–10 n.12.

60. There are concerns over fully automated analysis of small samples and mixtures because of sample consumption, reduced sensitivity relative to full-scale equipment, and the integrated software for interpretation. Discussions of the state of the art and research and administrative recommendations can be found in, for example, Erik Dalin et al., Biology Section, Nat'l Forensic Centre, Swedish Police Authority, Rapid DNA: A Summary of Available Rapid DNA Systems (2022), https://perma.cc/87XS-WQ9W; Douglas R. Hares et al., *Rapid DNA for Crime Scene Use: Enhancements and Data Needed to Consider Use on Forensic Evidence for State and National DNA Databasing—An Agreed Position Statement by ENFSI, SWGDAM and the Rapid DNA Crime Scene Technology Advancement Task Group*, 48 Forensic Sci. Int'l: Genetics 102349 (2020), https://doi.org/10.1016/j.fsigen.2020.102349; Non-CODIS Rapid DNA Best Practices/Outreach and Courtroom Considerations Task Group, Texas Forensic Sci. Comm'n, Non-CODIS Rapid DNA Considerations and Best Practices for Law Enforcement Use, Sept. 16, 2019, https://perma.cc/X3VV-RCUB.

61. *E.g.*, Belinda Martin et al., *Analysis of Rapid HIT Application to Touch DNA Samples*, 67 J. Forensic Sci. 1233 (2022), https://doi.org/10.1111/1556-4029.14964 ("not fit for the routine analysis of touch DNA samples in forensic casework"); Denise Ward et al., *Analysis of Mixed DNA Profiles from the RapidHIT™ ID Platform Using Probabilistic Genotyping Software STRmix™*, 58 Forensic Sci. Int'l: Genetics 102664 (2022), https://doi.org/10.1016/j.fsigen.2022.102664 ("good discrimination power [when STR data was analyzed with external software] but less than those produced via the standard laboratory workflow," as would be "expected . . . for the advantages of speed and portability").

Sequence-Specific Probes and Microarrays

Simple sequence variation, such as that for the GC locus (see section titled "Chromosomes" above), is conveniently detected using sequence-specific probes (defined in the sections on RFLP-VNTR and STR profiling). With GC typing, for example, probes for the three common alleles are attached at three separate spots on a membrane. Copies of the variable sequence region of the GC gene in the crime-scene sample are made with PCR. These copies (in the form of single strands) are poured onto the membrane. Whichever allele is present in a copy will cause it to stick to a corresponding, immobilized probe strand. A chemical "label" that catalyzes a color change at the spot where the trace-evidence allele binds to its probe can be attached when the copies are made. A colored spot showing that the allele is present thus should appear on the membrane at the location of the probes that correspond to this particular allele. If only one allele is present in the crime-scene DNA (because of homozygosity), there will be no color change at the spots where the other probes are located. If two alleles are present (heterozygosity), the corresponding two spots will change color.

This method was used before STR profiling, with only a few loci.[62] In that form, it lacked the discriminating power of STRs and VNTRs, but, being PCR-based, it was more sensitive, and easier and quicker to perform than RFLP-VNTR profiling. Admissibility followed.[63]

The approach can be miniaturized and automated by embedding probes for many loci on a chip.[64] Such a "microarray" for DNA-sequence detection usually consists of a two-dimensional grid of many thousands of microscopic spots on a silicon, glass, plastic, or paper surface. Each spot contains many copies of a probe with its own particular sequence, tethered to the surface at one end. A solution containing copies of single-stranded target DNA fragments[65] (with attached fluorescent tags or dyes) is washed over the microarray surface. As usual, the probes on the array bind to copies with the complementary sequence. The spots

62. Indeed, in what probably was the first criminal case in the United States in which DNA evidence was admitted, a single-locus test that suggested that the owners of a nursing home, although found guilty of negligent homicide for the starvation of an elderly patient, did not swap the organs of the victim with those of another cadaver in an alleged attempt to cover up the cause of death. Stephen Michaud, *DNA Detectives*, N.Y. Times Mag., Nov. 6, 1988, § 6, at 70, available at https://www.nytimes.com/1988/11/06/magazine/dna-detectives.html.

63. On the development and judicial reception of the "dot blot" or "reverse dot blot" tests and the "polymarker" system, see Kaye, *supra* note 1, at 180–87.

64. *See, e.g.,* Johannes Wöhrle et al., *Digital DNA Microarray Generation on Glass Substrates*, 10 Sci. Reps. 5770 (2020), https://doi.org/10.1038/s41598-020-62404-1.

65. Amplification (including isothermal methods that do not require the heating and cooling cycles of conventional PCR) of the target DNA can be built into the microarray. Roger Bumgarner, *Overview of DNA Microarrays: Types, Applications, and Their Future*, 101 Current Protocols in Molecular Biology 22–1 (2013), https://doi.org/10.1002/0471142727.mb2201s101.

that capture the targeted DNA are identified, indicating the presence of those sequences in the sample.

DNA microarrays have broad applications in biology and medicine.[66] Forensic applications include sequencing human mitochondrial DNA (see section titled "Mitochondrial DNA" below) and resolving individual genotypes within complex DNA mixtures,[67] but the most well known application comes from SNP chips with enough different probes to detect half-a-million or more different known SNPs distributed throughout the human genome.[68] After biotechnology companies produced these microarrays for genomics research and health applications,[69] direct-to-consumer genetic testing companies popularized them for ancestry and other "recreational genetics" activities. As more and more people interested in locating possibly unknown relatives elected to share certain genomic information in a public database, trawling the database for putative relatives of perpetrators of cold cases to infer the identity of the unknown source of the crime-scene DNA proved spectacularly successful. This procedure of investigative genetic genealogy is the subject of a later section by that name.

Sequencing and SNPs

As shown in Figure 1 (see section titled "The DNA Molecule" above), DNA contains a sequence of paired letters—the nucleotides A, T, G, and C—laid out

66. They permit large-scale population studies—for example, to determine how often individuals with a particular mutation develop breast cancer, or to identify the changes in gene sequences that are most often associated with particular diseases. Microarrays also are used in studies of variation in the number of copies of certain genes in different people's genomes (copy number variation). They are used to study the extent to which certain genes are turned on or off in cells and tissues. For this purpose, instead of isolating DNA from the samples, a transcript of the DNA (mRNA, *supra* section titled "Chromosomes") is isolated and measured. They provide clinical diagnostic tests for diseases. "Biosensor" microarrays detect pathogens and other targets.

67. Lev Voskoboinik et al., *SNP-Microarrays Can Accurately Identify the Presence of an Individual in Complex Forensic DNA Mixtures*, 16 Forensic Sci. Int'l: Genetics 208 (2015), https://doi.org/10.1016/j.fsigen.2015.01.009 (proof of concept).

68. Another proposed use is providing SNP data for better discerning the distinct contributors to DNA mixtures. Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLoS Genetics e1000167 (2008), https://doi.org/10.1371/journal.pgen.1000167. *But see* Rosemary Braun et al., *Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data*, 5 PLoS Genetics e1000668 (2009), https://doi.org/10.1371/journal.pgen.1000668; Peter M. Visscher & William G. Hill, *The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis*, 5 PLoS Genetics e1000628, https://doi.org/10.1371/journal.pgen.1000628.

69. One study of 3,000 Europeans used a commercial microarray with over half a million SNPs "to infer [the individuals'] geographic origin with surprising accuracy—often to within a few hundred kilometers." John Novembre et al., *Genes Mirror Geography Within Europe*, 456 Nature 98, 98 (2008), https://doi.org/10.1038/nature07331.

in a spiraling kind of line. Broadly, sequencing refers to ascertaining the precise order of the DNA letters along each strand.[70] The sequence-specific probes of the previous section could be said to sequence short segments of DNA in those instances in which they bind to the complementary sequence in the target DNA. But that is not how DNA sequencing machines work. They determine the order of base pairs one at a time, somewhat like spelling out the letters of a word before one recognizes it as a word. This sequencing is the subject of this section. It can be done for short stretches of DNA or, with enough effort and ingenuity, and in a roundabout way, for the whole genome, ultimately revealing the order of the base pairs at every gene, every STR, and every other kind of locus.

Sequencing methods

Sequencing technology is hardly new. In the 1970s, biochemists developed two PCR-related methods to sequence DNA.[71] For the next 30 years, researchers applied "Sanger sequencing" to decipher complete genes and, later, entire genomes of organisms. It is still used to sequence particular regions (up to a thousand or so base pairs). Basically, it uses PCR to synthesize a series of complementary DNA strands with color-coded nucleotides that indicate whether the nucleotide in the template strand is an A, T, G, or C.[72] The multinational, multiyear Human Genome Project, completed in the early 2000s, employed factory-like systems with hundreds of sequencing machines to Sanger sequence most of the 3.2 billion base pairs in the haploid genome from several anonymous

70. Because of the strict A-T, G-C pairing rule, it is not necessary to write two letters to designate a pair; the sequence of bases on one strand implies the sequence on the other.

71. Allan M. Maxam & Walter Gilbert, *A New Method for Sequencing DNA*, 74 Proc. Nat'l. Acad. Sci. 560 (1977), https://doi.org/10.1073/pnas.74.2.560; Frederick Sanger et al., *DNA Sequencing with Chain-terminating Inhibitors*, 74 Proc. Nat'l. Acad. Sci. 5463 (1977), https://doi.org/10.1073/pnas.74.12.5463.

72. As with ordinary PCR amplification, the target DNA is unzipped and copied—but the polymerase chain reaction stops prematurely because some "chain-terminating" nucleotides (the A, T, C, and G units) are added to the PCR mixture. The chemically modified nucleotides are missing the molecular "hook" that lets the next unit attach itself to the growing chain, and they have a fluorescent dye attached to them (with a different color for A, T, G, and C units). The chain reaction continues adding nucleotides until it happens to add a modified one. This process is repeated many times to generate fragments that extend to all possible stopping points. The smallest ones will consist of only one terminating unit after the primer. The next smallest will have two nucleotides, and so on, for every base pair in the original target DNA fragment. Capillary electrophoresis (*supra* section titled "STRs and Capillary Electrophoresis") separates the synthesized fragments from smallest to largest, with the results displayed as colored peaks in the electropherogram (also called a chromatogram). The order of the colored peaks gives the sequence of the base pairs in the target DNA. The whole process is an example of sequencing by synthesis.

individuals.[73] Despite the ultimate success, the sequencing efforts were technically demanding, expensive, time-consuming, and not suitable for regions on the chromosomes that are replete with repetitive DNA.[74]

In 2004, the National Human Genome Research Institute announced funding for research leading to the "$1,000 genome," which would permit sequencing an individual's genome for medical diagnosis and improved drug therapies. The goal was essentially met in 2014 thanks to methods known as next-generation sequencing (NGS) or massively parallel sequencing (MPS).[75] MPS methods simultaneously read millions of fragments and then use computer-intensive alignment programs to stitch the reads together to give the whole genome sequence. The genomes can be small (as with bacteria), large (as with humans), or larger still (as with some plants and other organisms). The methods also can be applied to selected parts of a genome, making them attractive for forensic and diagnostic purposes.

Commercially available sequencers implement the MPS approach in different ways.[76] One popular system is another form of sequencing by synthesis. First, to prepare the sample for sequencing, all the DNA is cut into many little pieces, or the targets are amplified as small pieces, creating a "library" of fragments. Then "adapters" are added to the ends of the pieces in this library. These adapters are like tiny handles that help immobilize the DNA to facilitate sequencing. Second, instead of running PCR in a tube, the DNA is loaded onto a glass "flow cell" that has millions of tiny wells on its surface. Each of these wells can grab an adapter to capture a single piece of DNA from the library. Third, the fragments in the wells are amplified with PCR, resulting in a cluster of identical single-stranded DNA in a given well. Fourth, chemically modified nucleotides bind to the DNA template strand. As in Sanger sequencing, each nucleotide contains a fluorescent tag, but the terminator is removed after the color at each well is detected, allowing the next base to bind. The reactions are repeated hundreds of times to trace the sequence of nucleotides in the growing chain. The cycle of one-pair-at-a-time synthesis occurs in millions of wells at once. Finally, bioinformatics software fits the jigsaw puzzle of millions of pieces together.[77]

73. Int'l Human Genome Sequencing Consortium, *Finishing the Euchromatic Sequence of the Human Genome*, 431 Nature 931 (2004), https://doi.org/10.1038/nature03001; Int'l Human Genome Sequencing Consortium: *Initial Sequencing and Analysis of the Human Genome*, 409 Nature 860 (2001), https://doi.org/10.1038/35057062; *cf.* J. Craig Venter et al., *The Sequence of the Human Genome*, 291 Science 1304 (2001), https://doi.org/10.1126/science.1058040.

74. More efficient methods have enabled researchers to fill in the gaps, correct errors, and release "a truly complete human reference genome." Sergey Nurk et al., *The Complete Sequence of a Human Genome*, 376 Science 44 (2022), https://doi.org/10.1126/science.abj6987.

75. Erwin L. van Dijk et al., *Ten Years of Next-Generation Sequencing Technology*, 30 Trends in Genetics 418 (2014), https://doi.org/10.1016/j.tig.2014.07.001.

76. Barton E. Slatko et al., *Overview of Next-Generation Sequencing Technologies*, 122 Current Protocols in Molecular Biology e59 (2018), https://doi.org/10.1002/cpmb.59.

77. With MPS, each base has to be read not just once, but at least several times in the overlapping segments to ensure accuracy in fitting the reads together.

The MPS sequencing-by-synthesis method described above sometimes is cited as an example of second-generation sequencing. Its reads are much shorter and more prone to error (per read) than Sanger sequencing,[78] but the speed and cost are enormous improvements for sequencing large genomes.

The sequencing part of the process (the steps starting with insertion of the library of DNA fragments into the flow cell) need not be applied to whole genomes. The library can consist of copies of short segments derived from any interesting sequences.[79] MPS thus supports targeted sequencing of many polymorphisms of forensic interest all at once.[80]

A second example of an MPS technology has been called third- or even fourth-generation sequencing. It produces long reads from a single DNA molecule without having to amplify it by PCR and to attach and illuminate fluorescent labels. Proteins embedded into a biological membrane create exquisitely small holes, called nanopores. (A human hair is on the order of 100,000 nanometers wide; a DNA molecule is 2–12 nanometers wide.) Either single- or double-stranded DNA can be moved through multiple nanopores of a flow cell at a constant rate for tens of thousands of nucleotides. As a DNA molecule is ratcheted through a pore, it interrupts a flow of electrical current to a sensor chip. The disruption produces a disturbance in the current that is characteristic of the type of nucleotide (A, T, G, or C) passing through. Software decodes the disturbance to determine the DNA sequence as the nucleotides flow by.

With such long-read sequencing, assembling the overlapping reads is easier, just as putting together a jigsaw puzzle with fewer, larger pieces is easier than fitting together numerous smaller ones. Moreover, large insertions or deletions and repetitive regions can be detected more easily. Although the longer reads can have a higher error rate, ongoing technological improvements are closing the accuracy gap with the sequencing-by-synthesis versions of MPS.[81]

78. Sanger sequencing platforms "offer well-trusted, up to 1-kbp long reads . . . and are still considered the gold standard for sequencing. They are routinely used for clinical gene tests or validations. However, their low-throughput limits them to single or small batches of targets." Adam Ameur et al., *Single-Molecule Sequencing: Towards Clinical Applications*, 37 Trends in Biotech. 72 (2019), https://doi.org/10.1016/j.tibtech.2018.07.013.

79. Panels of genes with variants that are known to promote the growth of cancer cells can be targeted to see which mutations are present in cancer patients. *See, e.g.,* Masayuki Nagahashiet et al., *Next Generation Sequencing-based Gene Panel Tests for the Management of Solid Tumors*, 110 Cancer Sci. 6 (2019), https://doi.org/10.1111/cas.13837.

80. Peter de Knijff, *From Next Generation Sequencing to Now Generation Sequencing in Forensics*, 38 Forensic Sci. Int'l: Genetics 175 (2019), https://doi.org/10.1016/j.fsigen.2018.10.017.

81. *E.g.*, Søren M. Karst et al., *High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or Pacbio Sequencing*, 18 Nature Methods 165 (2021), https://doi.org/10.1038/s41592-020-01041-y.

High-throughput sequencing technologies have demonstrated their usefulness in a wide range of research,[82] clinical,[83] and public health[84] applications. A famous example is the whole-genome sequencing of highly degraded ancient DNA (aDNA) of extinct species, including woolly mammoths, cave bears, Neanderthals, and Denisovans, as well as more modern human populations, from Vikings to Paleo-Inuit.[85]

The range of forensic applications

Of course, different applications within the broad category of massively parallel or next-generation sequencing can use different techniques, and particular forensic applications need to be validated to demonstrate that they are fit for their intended use. Research has shown that MPS can be used for the following:

- to increase the power of STR loci to distinguish among individuals by detecting sequence differences within the length-based alleles as measured by capillary electrophoresis;[86]

82. *E.g.*, Vivian Marx, *Method of the Year: Long-read Sequencing*, 20 Nature Methods 6 (2023), https://doi.org/10.1038/s41592-022-01730-w; Katia Nones & Ann-Marie Patch, *The Impact of Next Generation Sequencing in Cancer Research*, 12 Cancers (Basel) 2928 (2020), https://doi.org/10.3390/cancers12102928.

83. Elaine Hsu et al., *Rapid Pathogen Detection by Metagenomic Next-Generation Sequencing of Infected Body Fluids*, 27 Nature Med. 115 (2021), https://doi.org/10.1038/s41591-020-1105-z; F. Mosele et al., *Recommendations for the Use of Next-Generation Sequencing (NGS) for Patients with Metastatic Cancers: A Report from the ESMO Precision Medicine Working Group*, 31 Annals Oncology 1491 (2020), https://doi.org/10.1016/j.annonc.2020.07.014. *But see* Francois Balloux et al., *From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic*, 26 Trends in Microbiology 1035 (2018), https://doi.org/10.1016/j.tim.2018.08.004 (discussing obstacles to more routine use); Nagahashiet et al., *supra* note 79.

84. *E.g.*, Rowena A. Bull et al., *Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis*, 11 Nature Commun. 6272 (2020), https://doi.org/10.1038/s41467-020-20075-6; David F. Nieuwenhuijse et al., *Towards Reliable Whole Genome Sequencing for Outbreak Preparedness and Response*, 23 BMC Genomics 569 (2022), https://doi.org/10.1186/s12864-022-08749-5; Joshua Quick et al., *Real-time, Portable Genome Sequencing for Ebola Surveillance*, 530 Nature 228 (2016), https://doi.org/10.1038/nature16996. By sequencing entire bacterial genomes, researchers can rapidly differentiate organisms that have been genetically modified for biological warfare or terrorism from routine clinical and environmental strains. B. La Scola et al., *Rapid Comparative Genomic Analysis for Clinical Microbiology: The Francisella Tularensis Paradigm*, 18 Genome Res. 742 (2008), https://doi.org/10.1101/gr.071266.107.

85. Elizabeth D. Jones, Ancient DNA: The Making of a Celebrity Science (2022); Andrew Curry, *Ancient DNA Pioneer Svante Pääbo Wins Nobel*, 378 Science 12 (2022), https://doi.org/10.1126/science.adf1845.

86. For example, one allele might have nine copies of a GATA repeat (as in Figure 2), while another might have eight GATA repeats and one GTTA instead of a GATA. Their PCR products, being the same length, would migrate through the gel at the same speed, so they would show up as a single peak on an electropherogram. So would two alleles—from a different individual—that are

- to better resolve mixtures into the components coming from different contributors (as discussed in the section titled "Mixtures" below);
- to develop investigative leads based on biogeographic ancestry and physical characteristics such as eye color;[87]
- to ascertain the tissues that the DNA in crime-scene samples came from when the whole cells are not present in the samples;[88]
- to sequence mitochondrial DNA;[89] and even
- to distinguish between identical twins.[90]

Moreover, as explained in the next section, the ability of MPS to find point mutations (insertions, deletions, or substitutions of single base pairs) opens up the possibility of using well-chosen SNPs as a replacement for (or at least a supplement to) STR profiling for the purpose of individual identification, that is, of determining whose DNA is in a crime-scene sample.[91] Biotechnology companies have developed kits for preparing the libraries for all these purposes,[92] and in 2024, MPS results were held to be admissible for the first time in the United States.[93]

SNPs as identification loci

Single-nucleotide polymorphisms (SNPs) are the most abundant variations in the human genome. SNPs are shorter than STRs, making it possible to type degraded samples that regular STR profiling cannot handle because the

made up of nine GATA repeats. PCR-CE would not distinguish between these two individuals, but the more detailed sequence information could. The additional information also can improve mixture resolution (*see infra* section titled "Mixtures").

87. *See infra* section titled "Investigative DNA Analysis."

88. *See infra* section titled "DNA and RNA Typing of Bodily Fluids or Tissues."

89. *See infra* section titled "Mitochondrial DNA."

90. *See infra* section titled "Twins."

91. *See, e.g.,* Christopher P. Phillips et al., *A Compilation of Tri-allelic SNPs from 1000 Genomes and Use of the Most Polymorphic Loci for a Large-Scale Human Identification Panel*, 46 Forensic Sci. Int'l: Genetics 102232 (2020), https://doi.org/10.1016/j.fsigen.2020.102232.

92. Penelope R. Haddrill. *Developments in Forensic DNA Analysis*, 5 Emerging Topics Life Sci. 381 (2021), https://doi.org/10.1042/ETLS20200304.

93. People v. Chavez, No. BF187877A (Kern Cnty. Cal. Super. Ct. Jan. 22, 2024). Defendant was convicted of first-degree murders of two homeless women. In this unreported case, "[f]orensic evidence, including DNA and fingerprints, directly linked [the defendant] to both deaths." Office of the Dis. Atty., Press Release, Jan. 24, 2024, https://perma.cc/4FX8-3K6G. Admission followed an "extensive" pretrial hearing on results obtained with a "MiSeq FGx System." *Id.* This "instrument [can] interrogate up to 96 combined SNP and STR libraries in a single run" using a massively parallel sequencing-by-synthesis procedure. Verogen MiSeq FGx Sequencing System for Next-generation Sequencing of Forensic Genomics Libraries, https://perma.cc/LUF7-YMSP.

fragments are shorter than the STR alleles.[94] Indeed, even with undegraded samples, STR analysts must cope with phenomena such as stutter peaks that are near a true peak but might be confused with a peak from a different allele. SNP analysis would obviate this problem with electrophoretic-STR data.

On the other hand, a SNP locus has fewer alleles than an STR locus (usually only two). Hence, it takes more of them to achieve high levels of discrimination. MPS systems solve this problem because they can analyze many separate SNP loci at once. Data on how frequently major SNP alleles occur in various population groups have been collected, so the significance of a match in a multilocus profile can be assessed numerically.[95] As such, some scientists see STR profiling as outdated—but recommend including STRs in the sequencing process if only to avoid having to redo the DNA databases of STR profiles of convicted offenders and arrestees.

Recent work with SNPs for individual identification focuses on using more than one SNP at a time for a marker. Two or three SNPs can be chosen within a short segment of DNA (smaller than 300 base pairs, for example). Such short blocks of DNA are very likely to be inherited as a package from one parent, and MPS targeted to them can identify this "microhaplotype" in a single sequence run. A pair of SNPs has more possible types than a single SNP, so each microhaplotype is more informative than a single-SNP locus. Panels of microhaplotypes have been studied for individual identification and other forensic-science applications.[96] With a sufficient number of SNPs, the power to discriminate between two randomly selected individuals is superior to that of conventional STR profiling.[97]

94. Moreover, SNPs have advantages for forensic tests that do not make direct comparisons of samples of known (from a suspect, for example) and unknown (from a crime scene) origin. They tend to be specific to certain populations, so they can be used to infer ancestry. *See infra* section titled "Biogeographic Ancestry Testing." Most have far lower mutation rates, which is advantageous in parentage and other kinship testing.

95. *See infra* sections titled "Could a Close Relative Be the Source?" and "Frequencies, Probabilities, and Prejudice."

96. To reduce the possibility of conveying significant information about disease status or susceptibility, microhaplotype loci can be limited to SNPs located far from genes. Ones that are very close to certain genes would tend to be inherited along with the gene, potentially allowing them to be used as markers for disease-causing gene mutations.

97. *E.g.*, Kenneth K. Kidd et al., *Evaluating 130 Microhaplotypes Across a Global Set of 83 Populations*, 29 Forensic Sci. Int'l: Genetics 29 (2017), https://doi.org/10.1016/j.fsigen.2017.03.014; Aliye Kureshi et al., *Construction and Forensic Application of 20 Highly Polymorphic Microhaplotypes*, 7 Royal Soc'y Open Sci. 191937 (2020), https://doi.org/10.1098/rsos.191937; Jing-Bo Pang et al., *A 124-plex Microhaplotype Panel Based on Next-generation Sequencing Developed for Forensic Applications*, 10 Sci. Reps. 1945 (2020), https://doi.org/10.1038/s41598-020-58980-x; Andrew J. Pakstis et al., *The Population Genetics Characteristics of a 90 Locus Panel of Microhaplotypes*, 140 Hum. Genetics 1753 (2021), https://doi.org/10.1007/s00439-021-02382-0.

Summary

DNA contains the genetic information of an organism. In humans, most of the DNA is found in the cell nucleus, where it is organized into separate chromosomes. Each chromosome is like a book, and each cell has the same set of books of various sizes. There are two copies of each book (a "diploid" genome)—one copy from the father, one from the mother (the two "haploid" genomes). Thus, there are two copies of the book entitled "Chromosome One," two copies of "Chromosome Two," and so on. The text is mostly the same in a pair of books, but sexual reproduction results in important differences as well as inconsequential ones.

Genes are like paragraphs in the books. They encode different proteins and RNAs. Parts of the DNA text appear to have no coherent message, but some of the noncoding DNA affects the production of the gene products. Two individuals sometimes have different versions (alleles) of the same gene as well as variants of the text outside of (or interrupting) the genes. Some alleles result from the substitution of one letter for another. These are SNPs. Others come about from the insertion or deletion of single letters, and still others represent a kind of stuttering repetition of a string of extra letters. These are the VNTRs and STRs.[98] The locations within a chromosome where these interpersonal variations occur are called loci.

Historically, forensic DNA typing relied primarily on the length differences of a small number of VNTR and then a larger number of STR loci.[99] STR profiling depends on a chemical reaction for generating a great many copies of DNA segments (PCR amplification), followed by separation of the copies according to their length (capillary electrophoresis). We can refer to this well-established technology as PCR-CE for STRs, or even more compactly, as STR-CE profiling. Modern sequencing technology enables more and improved SNP-based loci for identity determinations and other forensic-science applications.

What Establishes That a Genetic System Is Valid for Identification?

Regardless of the kind of genetic system used for typing—STR, SNP, or still other polymorphisms—some general principles and questions can be applied to judge its validity.[100] First, the nature of the polymorphism should be well

98. In addition to the 23 pairs of "books" (chromosomes) in the cell nucleus, other scraps of DNA "text" reside in each of the mitochondria, the power plants of the cell. *See* section titled "Mitochondrial DNA" below.

99. For a short time, a few SNPs in protein-coding DNA also were used.

100. In addition to the suggestions on studying validity in this section, see section titled "Validation" below.

characterized. Is it a sequence polymorphism or a length polymorphism? Where is it situated within the genome? This information should be in the published literature or in archival genome databanks.

Second, the published scientific literature can be consulted to verify claims that a particular method of analysis can produce accurate profiles under various conditions. Ideally, these claims will rest on a detailed understanding of how the analytical procedure works—that is, of how measurements are made and how inferences are drawn from these measurements—as well as empirical studies of the results achieved with the methods under experimental conditions or in casework.[101] Although such validation studies have been conducted for all the systems described here, determining the point at which the validation of a particular system is sufficiently extensive and convincing to pass scientific muster may well require expert assistance.[102] The standards for validation that have been issued by private standards-developing organizations or expert groups often have only general requirements.[103]

Finally, the population genetics of the system should be characterized. As new systems are discovered, researchers typically analyze convenient collections of DNA samples from various human populations and publish studies of the relative frequencies of each allele in these population samples.[104] These studies give a measure of the extent of variability at the polymorphic locus in the various populations, and thus of the potential probative power of the marker for distinguishing among individuals.

At this point, the capability of the widely used PCR-CE procedures to ascertain STR genotypes accurately cannot be doubted.[105] But the technology has its limits, and a laboratory should be able to show that it is operating within the limits

101. *See id.*

102. *See* section titled "Relevant Expertise" above.

103. When it comes to specifying what level of performance an analytical method must be shown to possess, many standards are vague or vacuous. For instance, the FBI's Scientific Working Group on DNA Analysis Methods conflates accuracy with repeatability and reproducibility of measurements and cursorily asserts that these properties "should be evaluated." SWGDAM, Validation Guidelines for DNA Analysis Methods §§ 3.5.1 & 3.5.2 (Dec. 5, 2016). ANSI/ASB Standard 038 (2020), entitled "Standard for Internal Validation of Forensic DNA Analysis Methods," likewise contains no minimum levels for accuracy and reproducibility, and its "requirements" barely go beyond the command that "[t]he laboratory shall conduct internal validation studies on all forensic DNA analysis methodologies prior to implementation." More detailed standards for particular techniques are in progress.

104. On the populations that should be sampled, see section titled "Could an Unrelated Person Be the Source?" below.

105. *See, e.g.,* President's Council of Advisors on Sci. & Tech. (PCAST), Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Sept. 2016, at 75, https://perma.cc/R76Y-7VU ("single-source samples or simple mixtures of two individuals, such as from many rape kits, is an objective method that has been established to be foundationally valid"); *accord*, John M. Butler et al., DNA Mixture Interpretation: A NIST Scientific Foundation Review, at 3 & 7, June 2021, NISTIR 8351-DRAFT; *see also supra* Introduction.

for which it has been tested (or if not, to explain why the change in circumstances does not matter). If small samples are at issue, it will be important to ask what validity studies show about the performance of the system as sample size is reduced. If samples include DNA mixtures, the question of whether the validity studies show that the method works well enough with similarly complex samples will arise.[106]

As next-generation technologies are introduced in court, the same basic questions will need to be answered: What is the principle of the new technology? Is it simply an extension of existing technologies, or does it invoke entirely new concepts? Is the new technology used in research or clinical applications independent of forensic science? If so, are there grounds for concern that the new technology has limitations that might affect its application in the forensic sphere? Finally, what testing has been done to establish that the new technology is valid and reliable when used on forensic samples? For massively parallel sequencing and microarray technologies, the questions may be directed as well to the bioinformatics methods used to analyze and interpret the raw data. Obtaining answers to these questions will likely require input both from experts involved in technology development and application and from knowledgeable forensic experts.

Of course, the fact that scientists have shown that it is possible to extract DNA and to analyze it in a way that bears on the issue of identity does not mean that a particular laboratory has adopted a suitable protocol, is generally proficient in following it, and has implemented the procedure correctly in the case at bar. These more case-specific issues are considered next.

Sample Collection and Laboratory Performance

Sample Collection and Contamination

The primary determinants of whether DNA typing can be done on any particular sample are (1) the quantity of DNA present in the sample and (2) the extent to which it is degraded. Generally speaking, if a sufficient quantity of reasonable-quality DNA can be extracted from a crime-scene sample, no matter what the nature of the sample, DNA typing can be done. Thus, DNA typing has been performed on old blood stains, semen stains, vaginal swabs, hair, bone, bite marks, cigarette butts, drinking cups, guns, garments, urine, and fecal material. This section discusses what constitutes sufficient quantity and reasonable quality in the context of capillary electrophoresis of PCR-amplified short tandem repeats (STR-CE profiling), which has been the predominant method for forensic DNA identification for more than thirty years. Complications from

106. *See* section titled "Validation of PGS Programs" below.

contaminants and inhibitors also are discussed, and emerging alternatives to STR-CE profiling are noted. The treatment of samples that contain DNA from two or more contributors is discussed in the section titled "Mixtures" below.

Did the Sample Contain Enough DNA?

Amounts of DNA present in some typical kinds of samples vary from a trillionth or so of a gram (a picogram) for a hair shaft to several millionths of a gram (micrograms) for a postcoital vaginal swab.[107] Most PCR test protocols recommend samples on the order of 1 billionth of a gram (1 nanogram) for optimal yields. Normally, the number of amplification cycles for nuclear DNA is limited to 28, 29, or 30 to avoid detecting alleles from less than about 10 to 15 cell equivalents of DNA (66 to 100 picograms).[108]

Procedures for STR-CE typing of still smaller samples—down to a single cell's worth of nuclear DNA—have been studied. These have been shown to work, to some extent, with trace or contact DNA left on the surface of an object such as the steering wheel of a car.[109] Improved reaction chemistries, amplification to a greater number of PCR cycles,[110] and detection systems that provide greater sensitivity have made detection of PCR products associated with single

107. A combination of chemical and physical methods permits DNA to be isolated from the other chemicals in a specimen collected from a crime scene or an individual. The basics are fairly standard in academic research (*see generally* Akash Gautam, DNA and RNA Isolation Techniques for Non-Experts (2022)), but scientists in a forensic DNA laboratory often face a high volume of samples, a variety of sample types, limited quantities of biological material, DNA degradation, chemical inhibitors, and environmental contaminants that can interfere with later analysis. Hence, there is no single best DNA extraction method. The method of choice often depends on the biological sample concerned. Kevin Wai Yin Chong et al., *Recent Trends and Developments in Forensic DNA Extraction*, 3 WIREs Forensic Sci. e1395 (2020), https://doi.org/10.1002/wfs2.1395. To save time and to avoid the loss of DNA in the extraction, isolation, and quantitation process, direct PCR amplification techniques have been developed. *See, e.g.,* Sarah E. Cavanaugh & Abigail S. Bathrick, *Direct PCR Amplification of Forensic Touch and Other Challenging DNA Samples: A Review*, 32 Forensic Sci. Int'l: Genetics 40 (2018), https://doi.org/10.1016/j.fsigen.2017.10.005; Belinda Martin et al., *Comparison of Six Commercially Available STR Kits for Their Application to Touch DNA Using Direct PCR*, 4 Forensic Sci. Int'l: Reports 100243 (2021), https://doi.org/10.1016/j.fsir.2021.100243.

108. *But see* SWGDAM, Guidelines for STR Enhanced Detection Methods 2 (Oct. 6, 2014), https://perma.cc/NYF7-R9HH ("current . . . kits . . . entail 26 to 32 cycles of PCR amplification with 0.5 to 2 ng of DNA template").

109. On the cellular material deposited by contact with hands, see Julia Burrill et al., *Corneocyte Lysis and Fragmented DNA Considerations for the Cellular Component of Forensic Touch DNA*, 51 Forensic Sci. Int'l: Genetics 102428 (2021), https://doi.org/10.1016/j.fsigen.2020.102428.

110. STR-CE testing of LT-DNA samples with "extra" PCR cycles is unusual in the United States. The laboratory of the Office of the Chief Medical Examiner for New York City employed a 31-cycle protocol instead of the then-established 28 cycles that generally withstood objections at the trial court level. In 2020, however, the New York Court of Appeals held in *People v. Williams*, 147 N.E.3d 1131 (N.Y. 2020), that, in light of defendant's showing that there was a controversy

molecules of DNA commonplace in casework analysis.[111] In addition, laboratory studies have shown that PCR-related methods of whole-genome amplification can make copies of a single DNA molecule that then may yield usable STR profiles with STR-CE.[112]

But when PCR is used on such small samples of DNA, chance effects might result in one allele being amplified much more than another. Alleles can drop out, small peaks from unusual alleles at other loci can appear, stutter peaks tend to be larger in relation to their parent peak, and bits of extraneous DNA can contribute to the profile. Laboratories normally conduct experiments with samples at successively lower concentrations not only to determine an "analytical threshold" at which a true peak reliably can be distinguished from low-level "noise" but also to establish a higher "stochastic threshold" at which the sporadic effects become apparent.[113] Laboratory protocols for acquiring further data and interpreting electropherograms from samples that exhibit stochastic effects vary and are discussed in connection with the interpretation of mixed DNA samples and the coming of age of probabilistic genotyping software (see section titled "Mixtures" below).

Although there are tests to estimate the quantity of DNA in a sample, whether a particular sample contains enough human DNA to allow PCR-based typing cannot always be predicted accurately, and it is not scientifically necessary to apply an arbitrary threshold as a prerequisite for testing.[114] If a result is

among scientists on the use of the additional cycles, it was an abuse of discretion to deny his motion for a pretrial hearing on the general acceptance of the modified PCR procedure.

In response to criticism of the LT-DNA evidence in the unsuccessful prosecution of Sean Hoey for a terrorist bombing in Omagh, Northern Ireland, England briefly suspended its use of the Forensic Science Service's 34-cycle procedure. Duncan Campbell & Vikram Dodd, *Police Suspend Use of Discredited DNA Test After Omagh Acquittal*, Guardian, Dec. 22, 2007. After it reinstated the method (with an added quantification step), the English Court of Appeal opined that "Low Template DNA can be used to obtain profiles capable of reliable interpretation if the quantity of DNA that can be analysed is above the stochastic threshold [of] between 100 and 200 picograms." R. v. Reed, [2009] (CA Crim. Div.) EWCA Crim. 2698, ¶ 74; Denise Syndercombe Court, *Low Copy Number DNA: Where Next?*, 50 Med., Sci. & Law 55 (2010), https://doi.org/10.1258/msl.2010.010015.

111. David Moore et al., *A Comprehensive Study of Allele Drop-In over an Extended Period of Time*, 48 Forensic Sci. Int'l: Genetics 102332 (2020), https://doi.org/10.1016/j.fsigen.2020.102332.

112. Man Chen et al., *Comparison of CE- and MPS-based Analyses of Forensic Markers in a Single Cell After Whole Genome Amplification*, 45 Forensic Sci. Int'l: Genetics 102211 (2020) https://doi.org/10.1016/j.fsigen.2019.102211; Richard Jäger, *New Perspectives for Whole Genome Amplification in Forensic STR Analysis*, 23 Int'l J. Molecular Sci. 7090 (2022), https://doi.org/10.3390/ijms23137090; Xu Qiannan et al., *Evaluating the Effects of Whole Genome Amplification Strategies for Amplifying Trace DNA Using Capillary Electrophoresis and Massive Parallel Sequencing*, 56 Forensic Sci. Int'l: Genetics 102599 (2022) https://doi.org/10.1016/j.fsigen.2021.102599.

113. *See* John M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation 40–44 & 93–95 (2015).

114. DNA concentration levels such as 100 or 200 picograms have been suggested to differentiate a conventional DNA profile from a low-level target profile. *See supra* note 110. However, STR-CE technology has improved since these numbers were proposed (*see* Davis R.L. Watkins

obtained, and if the controls (samples of known DNA and blank samples) have behaved properly, then the sample had enough DNA.[115] This is so whether or not the label low-copy number (LCN) or low template (LT) applies.[116]

Was the Sample of Sufficient Quality?

The primary determinant of DNA quality for forensic analysis is the extent to which the long DNA molecules are intact. Within the cell nucleus, each molecule of DNA extends for millions of base pairs. Outside the cell, DNA spontaneously degrades into smaller fragments at a rate that depends on temperature, exposure to oxygen, and, most importantly, the presence of water.[117] In dry biological samples,

et al., *Revisiting Single Cell Analysis in Forensic Science*, 11 Sci. Reps. 7054 (2021), https://doi.org/10.1038/s41598-021-86271-6, & authorities cited), and "[t]hresholds are often difficult to apply in a meaningful way, for example, in a sample that comprises DNA from two or more individuals, the total quantifiable does not reflect the individual contributions." Forensic Science Regulator (U.K.), Guidance: The Interpretation of DNA Evidence (Including Low-Template DNA), FSR-G-202 § 5.2.4 (2020), https://perma.cc/TX2Y-5SXL.

115. Different views have been expressed on the desirability of performing replicate tests (splitting the sample into several PCR tubes) and on how to interpret the resulting profiles. The "consensus method" discards all possible alleles other than those that are detected in two or more of the resulting electropherograms. Simon Cowen et al., *An Investigation of the Robustness of the Consensus Method of Interpreting Low-Template DNA Profiles*, 5 Forensic Sci. Int'l: Genetics 400 (2011), https://doi.org/10.1016/j.fsigen.2010.08.010. As statistical methods for modeling the sources of variability have advanced, replicate testing has fallen out of favor. *See, e.g.,* David J. Balding, *Evaluation of Mixed-source, Low-template DNA Profiles in Forensic Science*, 110 Proc. Nat'l Acad. Sci. 12241 (2013), https://doi.org/10.1073/pnas.1219739110; Forensic Science Regulator, *supra* note 114, § 11.1.4. Replicate profiling of small samples continues to be discussed. *E.g.,* Forensic Science Regulator, *supra*, § 12; Simone Gittelson et al., *Low-template DNA: A Single DNA Analysis or Two Replicates?*, 264 Forensic Sci. Int'l 139 (2016), https://doi.org/10.1016/j.forsciint.2016.04.012; SWGDAM, *supra* note 108, § 4.1.

116. The phrase "low copy number" originally was coined for the Forensic Science Service's method that used 34 PCR cycles. It has caused confusion. *E.g.*, United States v. McCluskey, 954 F. Supp. 2d 1224, 1278 (D.N.M. 2013) (skirting the debate between the parties on whether LCN refers to modifications in the testing or to the quantity of DNA by focusing on "whether the Government's DNA results in this case are reliable and admissible" rather than "establish[ing] a definition of 'LCN testing'"); United States v. Davis. 602 F. Supp. 2d 658 (D. Md. 2009) (avoiding "making a finding with regard to the dueling definitions of LCN testing advocated by the parties"); State v. Bigger, 254 P.3d 1142, 1151–52 (Ariz. Ct. App. 2011) ("We need not determine the proper definition for low copy number, or whether the DNA testing used in this case is labeled properly as LCN."). The more generic term "low template" (LT) DNA analysis includes the use of longer injection times for CE and sample concentration methods. Forensic Science Regulator, *supra* note 114, § 5.2.1. SWGDAM, *supra* note 108, introduced a more complicated definition.

117. Other forms of chemical alteration to DNA are well studied, both for their intrinsic interest and because chemical changes in DNA are a contributing factor in the development of cancers in living cells. Some forms of DNA modification, such as those produced by exposure to ultraviolet radiation, inhibit the amplification step in PCR-based tests, whereas other chemical

protected from air and not exposed to temperature extremes, DNA degrades very slowly. STR testing has proved effective with old and badly degraded material such as the remains of the Tsar Nicholas family (buried in 1918 and recovered in 1991).[118]

The extent to which degradation affects a PCR-based test depends on the size of the DNA segment to be amplified. For example, in a sample in which the bulk of the DNA has been degraded to fragments well under 1,000 base pairs in length, it may be possible to amplify a 100-base-pair sequence, but not a 1,000-base-pair target. Consequently, the shorter alleles may be detected in a highly degraded sample, but the larger ones may be missed. Inasmuch as the size differences among STR alleles at a locus are quite small (typically no more than 50 base pairs), "locus dropout" is more common than the dropout of only one allele at a heterozygous locus.[119]

DNA can be exposed to a great variety of environmental insults without any effect on its capacity to be typed correctly. Exposure studies have shown that contact with a variety of surfaces, both clean and dirty, and with gasoline, motor oil, acids, and alkalis either have no effect on DNA typing or, at worst, render the DNA untypable.[120]

Although contamination with microbes generally does little more than degrade the human DNA, other problems sometimes can occur. Therefore, the validation of DNA typing systems should include tests for interference with a variety of microbes to see if artifacts occur. If artifacts are observed, then control tests should be applied to distinguish between the artifactual and the true results.

modifications appear to have no effect. Cydne L. Holt et al., *TWGDAM Validation of AmpFlSTR PCR Amplification Kits for Forensic DNA Casework*, 47 J. Forensic Sci. 66 (2002), https://doi.org/10.1520/jfs15206j; George F. Sensabaugh & Cecilia von Beroldingen, *The Polymerase Chain Reaction: Application to the Analysis of Biological Evidence*, *in* Forensic DNA Technology 63 (Mark A. Farley & James J. Harrington eds., 1991).

118. Peter Gill et al., *Identification of the Remains of the Romanov Family by DNA Analysis*, 6 Nature Genetics 130 (1994), https://doi.org/10.1038/ng0294-130.

119. In particular, in cases involving severe degradation, loci yielding products greater than 200 base pairs may not be detected. Holt et al., *supra* note 117. Special primers and very short STRs give better results with extremely degraded samples. *See* Michael D. Coble & John M. Butler, *Characterization of New MiniSTR Loci to Aid Analysis of Degraded DNA*, 50 J. Forensic Sci. 43 (2005), https://doi.org/10.1520/jfs2004216. Sequencing methods can be even more effective for such samples. *See supra* section titled "SNPs as identification loci."

120. Holt et al., *supra* note 117. Most of the effects of environmental insult readily can be accounted for in terms of basic DNA chemistry. For example, some agents produce degradation or damaging chemical modifications. Other environmental contaminants inhibit restriction enzymes or PCR. (This effect sometimes can be reversed by cleaning the DNA extract to remove the inhibitor.) But environmental insult does not result in the selective loss of an allele at a locus or in the creation of a new allele at that locus.

Laboratory Performance

What Quality Control and Assurance Measures Are in Place?

DNA profiling of suitable samples is valid and reliable when properly conducted, but confidence in a particular result also depends on the quality control and quality assurance procedures in the laboratory. Quality control refers to measures to help ensure that DNA-typing results and their interpretation meet a specified standard of quality. Quality assurance refers to monitoring, verifying, and documenting laboratory performance. A quality assurance program helps demonstrate that a laboratory is meeting its quality control objectives.[121]

Professional bodies within forensic science have described general procedures for quality assurance. Guidelines for DNA analysis have been prepared by FBI-appointed groups (the current incarnation is known as the Scientific Working Group on DNA Analysis Methods [SWGDAM]),[122] the federally supported Organization of Scientific Area Committees for Forensic Science (OSAC),[123] and private Standards Developing Organizations (SDOs).[124]

121. For a review of the history of quality assurance in forensic DNA testing, see Joseph L. Peterson et al., *The Feasibility of External Blind DNA Proficiency Testing. I. Background and Findings*, 48 J. Forensic Sci. 21, 22 (2003), https://doi.org/10.4324/9780203380826.

122. The FBI established the Technical Working Group on DNA Analysis Methods (TWGDAM) in 1988 to develop standards. The DNA Identification Act of 1994, 34 U.S.C. §§ 12591(a)–(c), 14131(a) & (c), created a DNA Advisory Board (DAB) to assist in promulgating quality assurance standards, but the legislation allowed the DAB to expire after five years (unless extended by the director of the FBI). 34 U.S.C. § 12591(b)(4). TWGDAM functioned under the DAB, 34 U.S.C. § 12591(a)(4), and was renamed the Scientific Working Group on DNA Analysis Methods (SWGDAM) in 1999. When the FBI allowed DAB to expire, SWGDAM replaced it. *See* Norah Rudin & Keith Inman, An Introduction to Forensic DNA Analysis 180 (2d ed. 2002); Paul C. Giannelli, *Regulating Crime Laboratories: The Impact of DNA Evidence*, 15 J.L. & Pol'y 59, 82–83 (2007). SWGDAM characterizes the FBI quality-assurance standards as establishing minimum requirements that it supplements with more detailed, noncompulsory guidelines. SWGDAM, Frequently Asked Questions, https://perma.cc/Z9AR-FG6A, last visited Sept. 9, 2025.

123. The Department of Commerce's National Institute of Standards and Technology (NIST) supports OSAC's standards-writing efforts and maintains "a repository of selected published and proposed standards for forensic science . . . to promote valid, reliable and reproducible forensic results." NIST, OSAC Registry (updated Dec. 22, 2023), https://perma.cc/M9LS-96UN. The OSAC Registry maintains that "the standards on this Registry have undergone a technical and quality review process that actively encourages feedback from forensic science practitioners, research scientists, human factors experts, statisticians, legal experts, and the public." *Id.* The standards are voluntary, and some have been criticized as vague or trivial. Geoffrey Stewart Morrison et al., *Vacuous Standards—Subversion of the OSAC Standards-development Process*, 2 Forensic Sci. Int'l: Synergy 206 (2020), https://doi.org/10.1016/j.fsisyn.2020.06.005. The rigor of the actual review process, particularly as it involves independent scientific and statistical review, also has been questioned. *See* Kaye et al., *supra* note 5, § 12.5.1, at 680–82; Maneka Sinha, *Radically Reimagining Forensic Evidence*, 73 Ala. L. Rev. 879 (2022).

124. The major SDO that publishes voluntary standards for forensic DNA analysis is the Academy Standards Board (ASB) of the American Academy of Forensic Sciences. AAFS, Academy

A small number of states require forensic DNA laboratories to be accredited,[125] and federal law requires accreditation or other safeguards for laboratories that receive certain federal funds[126] or participate in the national DNA database system.[127]

Documentation

Quality assurance guidelines normally call for laboratories to document laboratory organization and management, personnel qualifications and training, facilities, evidence control procedures, validation of methods and procedures, analytical procedures, equipment calibration and maintenance, standards for case documentation and report writing, procedures for reviewing case files and testimony, proficiency testing, corrective actions, audits, safety programs, and review of subcontractors.

Of course, maintaining documentation and records alone does not guarantee the correctness of results obtained in any particular case. Measurement error can be estimated but not eliminated, and process errors in analysis or

Standards Board (2022), https://perma.cc/Y8N4-A4FU. Inasmuch as the dominant voice in both OSAC and the SDOs that promulgate forensic-science standards is the practicing forensic-science community, a consensus standard does not necessarily reflect a consensus of experts in the broader scientific community. Within OSAC and the ASB, agreement of two-thirds or more of the eligible voters constitutes a consensus. The votes are not published.

125. Joseph Peterson & Matthew Hickman, *Forensic Science Practice in the United States*, *in* The Global Practice of Forensic Science 301, 331 (Douglas H. Ubelaker ed., 2015). New York was the first state to impose this requirement. N.Y. Exec. Law § 995-b (McKinney 2006) (requiring accreditation by the state Forensic Science Commission). Only one state, Texas, requires its forensic analysts to be licensed. Texas Forensic Sci. Comm'n, Forensic Analyst Licensing Program, https://perma.cc/6UDT-NU3W, last visited Sept. 9, 2025.

126. The Justice for All Act, enacted in 2004, required DNA labs to be accredited within two years "by a nonprofit professional association of persons actively involved in forensic science that is nationally recognized within the forensic science community" and to "undergo external audits, not less than once every 2 years, that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation." 34 U.S.C. § 12592(b)(2)(A). The 2004 Act also requires applicants for federal funds for forensic laboratories to certify that the laboratories use "generally accepted laboratory practices and procedures, established by accrediting organizations or appropriate certifying bodies," 34 U.S.C. § 10562(2), and that "a government entity exists and an appropriate process is in place to conduct independent external investigations into allegations of serious negligence or misconduct substantially affecting the integrity of the forensic results committed by employees or contractors of any forensic laboratory system, medical examiner's office, coroner's office, law enforcement storage facility, or medical facility in the State that will receive a portion of the grant amount." *Id.* § 10562(4).

127. *See* 34 U.S.C. § 12592(b)(2) (requiring that records in the database come from laboratories that "have been accredited by a nonprofit professional association . . . and . . . undergo external audits, not less than once every 2 years [and] that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation").

interpretation can occur as a result of a deviation from an established procedure, an analyst misjudgment, or an accident. Although administrative and technical review procedures within a laboratory should be designed to detect errors before a report is issued, it is always possible that some incorrect result will slip through.[128] Accordingly, determination that a laboratory maintains a strong quality-assurance program, either on paper or in practice, does not guarantee accuracy in all cases.[129]

Validation

The validation of procedures is central to quality assurance. Developmental validation is undertaken to determine the applicability of a new test or equipment to crime-scene samples; it defines conditions that give accurate results and identifies the limitations of the procedure.[130] For example, a new genetic locus being considered for use in forensic analysis will be tested in both fresh samples and in samples typical of those found at crime scenes. The validation would include samples originating from different tissues, samples containing degraded DNA, samples contaminated with microbes, samples containing DNA mixtures, and so on. A valid typing method usually will report a genotype that is in a test sample and usually will not report a genotype that is not in the test sample.[131] Developmental validation of a new set of loci also includes the generation of population databases and the testing of alleles for statistical independence. Developmental validation normally results in publication in the scientific literature and reflects scientific norms for theoretical and

128. *See* U.S. Dep't of Just., Office of the Inspector Gen., The FBI DNA Laboratory: A Review of Protocol and Practice Vulnerability (2004), available at https://oig.justice.gov/reports/fbi-dna-laboratory-review-protocol-and-practice-vulnerabilities.

129. The District of Columbia's independent, state-of-the-art Department of Forensic Sciences twice lost its accreditation—and each of its first two directors—following complaints from prosecutors, first about DNA mixture analyses and later about firearms-toolmark examinations. *See* Keith L. Alexander & Julie Zauzmer, *Director of D.C.'s Embattled DNA Lab Resigns After Suspension of Testing*, Wash. Post, May 1, 2015, at B1, https://perma.cc/Z4FM-EWWB; Editorial, *The D.C. Crime Lab Is in Trouble—Again*, Wash. Post, June 4, 2021, at A20.

130. *See supra* section titled "What Establishes That a Genetic System Is Valid for Identification?"

131. No test is *always* correct. For a set of test samples, all of which contain a particular genotype, a test might be positive for that genotype in, say, 96% of them. This proportion is the observed "sensitivity" in the test set. For a set of test samples, none of which contain the genotype, the test might be negative for that genotype in, say, 99% of them. This proportion is the observed "specificity" for that genotype in the test set. A perfect test would have 100% sensitivity and 100% specificity in all properly conducted experiments. For further discussion of the validity of binary (present-or-absent) testing, see Kaye & Stern, *supra* note 12.

empirical support of claims for methods,[132] but the empirical aspects for validating a new procedure can be accomplished in multiple laboratories well ahead of publication.[133]

So-called internal validation, on the other hand, involves the capacity of a specific laboratory to analyze the new loci or use the new methodology or equipment.[134] The laboratory should verify that it can reliably perform an established procedure that already has undergone developmental validation.[135] In particular, before adopting a new procedure, the laboratory should verify that its analysts can obtain correct results with samples that are representative of casework.

Proficiency testing

Proficiency testing in forensic genetic testing is designed to ascertain whether an analyst can correctly determine genetic types in a sample whose origin is unknown to the analyst but is known to a tester. Proficiency is demonstrated by making correct determinations in repeated trials. The laboratory also can be tested to verify that it correctly computes random-match probabilities or likelihood ratios.

An internal proficiency trial is devised and conducted within a laboratory.[136] One person in the laboratory prepares the sample and then administers the test to another person in the laboratory. In an external trial, the test sample originates

132. As such, validation is not easily reduced to standards defining the research that needs to be undertaken, and validation standards often seem open textured. *See supra* note 103.

133. SWGDAM "encourage[s]" but does not require publication of developmental validity studies in scientific venues. SWGDAM, *supra* note 103, § 2.2.1.2. Whether the validity required for admission of scientific evidence can be established in the absence of a robust set of scientific publications is a legal rather than a strictly scientific question. However, it has been argued that scientific norms demand multiple studies with confirmation coming from separate research groups rather than only from the developer of a method. President's Council of Advisors on Sci. & Tech., *supra* note 105, at 80.

134. This line between developmental and internal validation is not always clear. Successful internal validation efforts add to the body of knowledge demonstrating that an analytical method performs as it should and thus enhances or extends what might be seen as a weak but encouraging previous validation. Furthermore, "internal validation" sometimes designates finding information needed to fine-tune a method. That function is better referred to as calibration.

135. Validation builds on the accumulated body of knowledge and experience. Thus, some aspects of validation testing need be repeated only to the extent required to verify that previously established principles apply.

136. Some forensic-science organizations do not use the term "proficiency testing" for tests of examiner performance developed or conducted within a single laboratory. In their view, proficiency testing requires interlaboratory comparisons. OSAC, OSAC Preferred Terms 2 (Aug. 2022), https://perma.cc/KJC4-6HWE.

from outside the laboratory—from another laboratory, a commercial vendor, or a regulatory agency. In a declared (or open) proficiency trial, the analyst knows the sample is a proficiency sample. The DNA Identification Act of 1994 required proficiency testing for analysts in the FBI as well as those in laboratories participating in the national database or receiving federal funding,[137] but these matters are now left to the discretion of the director of the FBI.[138] The standards of accrediting bodies typically call for periodic open, external proficiency testing.[139]

In a blind or, more properly, "full blind" trial, the sample is submitted so that the analyst does not recognize it as a proficiency sample. A full-blind trial provides a better indication of proficiency because it ensures that the analyst will not give the trial sample any special attention, and it tests more steps in the laboratory's processing of samples. However, full-blind proficiency trials entail considerably more organizational effort and expense than open proficiency trials. Obviously, the "evidence" samples prepared for the trial have to be sufficiently realistic that the laboratory does not suspect the legitimacy of the submission. A police agency and prosecutor's office have to submit the "evidence" and respond to laboratory inquiries with information about the "case." Finally, the genetic profile from a proficiency test must not be entered into regional and national databases. Consequently, although some forensic DNA laboratories participate in full-blind testing, they are not required to do so.[140]

137. Pub. L. 103–22 § 210304(b)(2) (codified as amended 34 U.S.C. § 12592(b)(2)(A)(ii)) (requiring external proficiency testing of laboratories for participation in the national database); *id*. § 210305(a)(1)(A) (same for FBI examiners).

138. 34 U.S.C. § 12592(b)(2)(A)(ii) (requiring "external audits, not less than once every 2 years, that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation"). The standards require periodic external proficiency testing. FBI, Quality Assurance Standards for Forensic DNA Testing Laboratories § 13.1 (2020), https://le.fbi.gov/file-repository/forensic-qas-070120.pdf.

139. *See* Peterson et al., *supra* note 121, at 24 (describing the ASCL-LAB standards). Certification by the American Board of Criminalistics as a specialist in forensic-biology DNA analysis requires one proficiency trial per year. Accredited laboratories must maintain records documenting compliance with required proficiency-test standards.

140. However, laboratory management can blind a particular analyst within the laboratory by injecting disguised "cases" into the usual flow of cases. Section 210303(c) of the DNA Identification Act of 1994, Pub. L. 103–22, 108 Stat. 2069, required the director of the National Institute of Justice to report to Congress on the feasibility of establishing an external blind-proficiency-testing program for DNA laboratories. A National Forensic DNA Review Panel advised the director that "blind proficiency testing is possible, but fraught with problems" of the kind listed above. Peterson et al., *supra* note 121, at 30. It "recommended that a blind proficiency testing program be deferred for now until it is more clear how well implementation of the first two recommendations [the promulgation of guidelines for accreditation, quality assurance, and external audits of casework] are serving the same purposes as blind proficiency testing." *Id*.

How Are Samples Created and Handled?

Sample mishandling, mislabeling, or contamination, whether in the field or in the laboratory, is more likely to compromise a DNA analysis than is an error in genetic typing. For example, a sample mix-up due to mislabeling reference blood samples taken at the hospital could lead to an incorrect association of crime-scene samples to a reference individual or to incorrect exclusions. Similarly, packaging two items with wet blood stains into the same bag could result in a transfer of stains between the items, rendering it difficult or impossible to determine whose blood was originally on each item. Contamination in the laboratory may result in artifactual typing results or in the incorrect attribution of a DNA profile to an individual or to an item of evidence. Procedures should be prescribed and implemented to guard against such error.

Mislabeling or mishandling can occur when biological material is collected in the field, when it is transferred to the laboratory, when it is in the analysis stream in the laboratory, when the analytical results are recorded, or when the recorded results are transcribed into a report. Mislabeling and mishandling can happen with any kind of physical evidence and are of great concern in all fields of forensic science. Checkpoints should be established to detect mislabeling and mishandling along the line of evidence flow. Investigative agencies should have guidelines for evidence collection and labeling so that a chain of custody is maintained. Similarly, there should be guidelines, produced with input from the laboratory, for handling biological evidence in the field.

Professional guidelines and recommendations require documented procedures to ensure sample integrity and to avoid sample mix-ups, labeling errors, recording errors, and the like. They also mandate case review to identify inadvertent errors before a final report is released. Finally, laboratories must retain, when feasible, portions of the crime-scene samples and extracts to allow reanalysis.[141] However, retention is not always possible. For example, retention of original items is not to be expected when the items are large or immobile (for example, a wall or sidewalk). In such situations, a swabbing, scraping, or cutting of the stain from the item would typically be collected and retained. There also are situations where the sample is so small that it will be consumed in the analysis.

141. *See* FBI, *supra* note 138, § 7.4.1. Furthermore, failure to preserve potentially exculpatory evidence has been treated as a denial of due process and grounds for suppression. People v. Nation, 604 P.2d 1051, 1054–55 (Cal. 1980). In *Arizona v. Youngblood*, 488 U.S. 51 (1988), however, the Supreme Court held that a police agency's failure to preserve evidence not known to be exculpatory does not constitute a denial of due process unless "bad faith" can be shown. Ironically, DNA testing that was not available at Youngblood's trial established that he had been falsely convicted. Maurice Possley, *DNA Exonerates Inmate Who Lost Key Test Case: Prosecutors Ruined Evidence in Original Trial*, Chi. Trib., Aug. 10, 2000, at 6, https://perma.cc/F78E-E8K7.

Assuming that appropriate chain-of-custody and evidence-handling protocols are in place, the critical question is whether there are deviations in a particular case. Answering this question may require a review of the total case documentation as well as the laboratory findings. In addition, when retesting original evidence items or the material extracted from them is possible, it can be used to guard against error from mislabeling and mishandling. Should mislabeling or mishandling have occurred, reanalysis of the original sample and the intermediate extracts may detect not only the fact of the error but also the point at which it occurred.[142]

Contamination describes any situation in which foreign material is mixed with a sample of DNA.[143] As noted in the section titled "Was the Sample of Sufficient Quality?" above, contamination by nonbiological materials, such as gasoline or grit, can cause test failures, but it is not a source of genetic typing errors. Similarly, contamination with nonhuman biological materials, such as bacteria, fungi, or plant materials, is generally not a problem. These contaminants may accelerate DNA degradation, but they do not generate spurious human genetic types.

The contamination of greatest concern results from the addition of human DNA. This sort of contamination can occur in three ways. First, the crime-scene samples, by their nature, may contain a mixture of fluids or tissues from different individuals. Examples include vaginal swabs collected as sexual-assault evidence and bloodstain evidence from scenes where several individuals shed blood. The analysis of such mixtures is the subject of the "Mixtures" section below.

Second, the crime-scene samples may be inadvertently contaminated in the course of sample handling in the field or in the laboratory. "Elimination databases" of DNA profiles from laboratory personnel as well as police officers and emergency responders who are present at crime scenes can be used to see

142. Of course, retesting cannot correct all errors that result from mishandling of samples, but it is possible in some cases to detect mislabeling at the point of sample collection if the genetic-typing results on a particular sample are inconsistent with an otherwise consistent reconstruction of events. For example, a mislabeling of husband and wife samples in a paternity case might result in an apparent maternal exclusion, a very unlikely event. The possibility of mislabeling could be confirmed by testing the samples for gender and ultimately verified by taking new samples from each party under better-controlled conditions.

143. *Cf.* Forensic Science Regulator (U.K.), Forensic Science Regulator Guidance: Contamination Controls—Scene of Crime § 1.1.1, at 4 (2023) (FSR-GUI-0016), https://perma.cc/Q4D2-MGJU ("For the purposes of this guidance, contamination is defined as 'the undesirable introduction of DNA, or biological material containing DNA, to an item/exhibit recovered from an incident scene which is to be recovered/analysed and before a controlled forensic process is started'. This is distinct from the adventitious transfer of biological material to an exhibit that can also occur, usually prior to the exhibit or sample being recovered and before investigative agencies have intervened, this is often referred to as 'background DNA'."); Forensic Science Regulator (U.K.), Forensic Science Regulator Guidance: DNA Contamination Controls—Laboratory § 1.1.1, at 5 (2023), FSR-GUI-0018, https://perma.cc/GU96-M769 (same).

if DNA from these individuals might be in the recovered samples.[144] Inadvertent contamination of crime-scene DNA with DNA from a reference sample could lead to a false inclusion. Likewise, with low-template DNA samples, secondary transfer—in which some DNA molecules from one person are present for a time on another person and then are deposited in the crime-scene sample—could be falsely incriminating.[145] Research into mechanisms of DNA transfer and DNA- and RNA-based testing that might help clarify the origin of the trace DNA is discussed in the section titled "DNA and RNA Typing of Bodily Fluids or Tissues" below.

Third, carryover contamination in PCR-based typing can occur if the amplification products of one typing reaction are carried over into the reaction mix for a subsequent PCR reaction. If the carryover products are present in sufficient quantity, they could be preferentially amplified over the target DNA. To protect against carryover contamination, the primary strategy is to keep PCR products away from sample materials and test reagents by having separate work areas for pre-PCR and post-PCR sample handling, by preparing samples in controlled-airflow biological safety hoods, by using dedicated equipment (such as pipettes) for each of the various stages of sample analysis, by decontaminating work areas after use (usually by wiping down or by irradiating with ultraviolet light), and by having a one-way flow of sample from the pre-PCR

144. Martine Lapointe et al., *Leading-edge Forensic DNA Analyses and the Necessity of Including Crime Scene Investigators, Police Officers and Technicians in a DNA Elimination Database*, 19 Forensic Sci. Int'l: Genetics 50 (2015), https://doi.org/10.1016/j.fsigen.2015.06.002; OSAC, Best Practice Recommendations for the Management and Use of Quality Assurance DNA Elimination Databases in Forensic DNA Analysis (Apr. 6, 2021) (2020-N-0007). However, the Genetic Information and Nondiscrimination Act of 2008, 42 U.S.C. § 2000ff et seq., prohibits employers from requesting "genetic information" from employees. An exception, in § 202(b)(6) of the act, applies "where the employer conducts DNA analysis for law enforcement purposes as a forensic laboratory or for purposes of human remains identification, and requests or requires genetic information of such employer's employees . . . for quality control to detect sample contamination." 42 U.S.C. § 2000ff-1(b)(6). Non-laboratory police employees have claimed that because they fall outside this exception, the statute protects them from being required to provide elimination samples. In response, it can be argued that the statute's expansive definition of "genetic test" should be read in light of its main purpose as an anti-discrimination law concerned with health-related genetic tests. David H. Kaye, *GINA's Genotypes*, 108 Mich. L. Rev. First Impressions 51 (2010). This intent- or purpose-based approach to the statute was rejected in *Lowe v. Atlas Logistics Group Retail Services (Atlanta)*, 102 F. Supp. 3d 1360 (N.D. Ga. 2015) (construing GINA on the basis of its plain text as prohibiting an employer's request for a DNA sample for STR profiling to find "a mystery employee [who was] habitually defecating in one of its warehouses"); *see also* David H. Kaye, *EEOC Stays Mum on GINA*, Forensic Sci., Stat. & L. (Jan. 22, 2013), https://perma.cc/QS9Z-XAYC.

145. *See* Peter Gill, Misleading DNA Evidence: A Guide for Scientists, Judges, and Lawyers (2014).

to post-PCR work areas.[146] Additional protocols are used to detect carryover contamination.[147]

In the end, whether a laboratory has conducted proper tests depends both on the general standard of practice and on the questions posed in the particular case. There is no universal checklist, but the selection of tests and the adherence to the correct test procedures can be reviewed by experts and by reference to professional standards.

Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing

What Constitutes a Match? An Exclusion?

The results of DNA testing can be presented in various ways. With discrete allele systems, such as STRs, it is natural—but not essential—to speak of "matching" and "nonmatching" profiles. If the profile obtained from the single-source biological sample taken from the crime scene or the victim (the trace-evidence sample) clearly differs from the DNA in a sample known to come from a particular individual (the reference sample), that individual is excluded as a possible source. At the other extreme, two multilocus genotypes can be clearly identical. In these cases, the DNA evidence typically is quite incriminating,[148] but even when two samples have the same genotype, there is a chance that the trace-evidence sample came not from the defendant, but from another individual who has the same genotype. As indicated in the section above titled "A Brief History

146. SWGDAM, Contamination Prevention and Detection Guidelines for Forensic DNA Laboratories (Jan. 12, 2017), https://perma.cc/R2SD-BJ8A.

147. Standard protocols include the amplification of blank control samples—those to which no DNA has been added. If carryover contaminants have found their way into the reagents or sample tubes, these will be detected as amplification products. Outbreaks of carryover contamination can also be recognized by monitoring test results. Detection of an unexpected and persistent genetic profile in different samples indicates a contamination problem. When contamination outbreaks are detected, corrective actions should be taken, and both the outbreak and the corrective action should be documented. Human Forensic Biology Subcomm., Org. of Sci. Area Committees for Forensic Sci., Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events (OSAC 2020-S-0004, Ver. 2.0, May 2021).

148. Whether being the source of the forensic sample is incriminating and whether someone else being the source is exculpatory depends on the circumstances. For example, a suspect who might have committed the offense without leaving the trace-evidence sample still could be guilty. In a rape case with several rapists, a semen stain could fail to incriminate one assailant because insufficient semen from that individual is present in the sample.

of DNA Evidence," this complication has produced extensive arguments over the statistical procedures for assessing this possibility.

Some cases lie between the poles of a clear inclusion or exclusion. Since the earliest days of RFLP-VNTR testing, concerns have been expressed about subjective aspects of procedures that leave room for "observer effects"[149] in interpreting data.[150] When the trace-evidence sample is small and extremely degraded, STR profiling can be afflicted with allelic drop-in, drop-out, and other complications.[151] Analysts operating within the framework of reporting that alleles are categorically present or absent then must make judgments as to whether true peaks are missing and whether spurious peaks are present.[152] Experts then might disagree about whether a suspect is included or excluded—or whether any conclusion can be drawn.[153]

What Hypotheses Can Be Formulated About the Source?

In DNA identification cases, the laboratory analyst reports that the sample of DNA from the crime scene and a sample from the defendant have the same genotype. There could be an innocent explanation for the defendant's DNA being at the crime scene,[154] but the matching DNA might not even be the defendant's. One

149. *See generally* D. Michael Risinger et al., *The* Daubert/Kumho *Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 Calif. L. Rev. 1 (2002).

150. *E.g.,* William C. Thompson & Simon Ford, *The Meaning of a Match: Sources of Ambiguity in the Interpretation of DNA Prints*, *in* Forensic DNA Technology (Mark J. Farley & James J. Harrington eds., 1990).

151. *See supra* section titled "Sample Collection and Contamination."

152. Commentators have proposed that the analyst determine the profile of a trace-evidence sample before knowing the profile of any suspects. *E.g.*, Itiel E. Dror & Jeff Kukucka, *Linear Sequential Unmasking-Expanded (LSU-E): A General Approach for Improving Decision Making As Well As Minimizing Noise and Bias*, 3 Forensic Sci. Int'l: Synergy 100161 (2021), https://doi.org/10.1016/j.fsisyn.2021.100161; Dan E. Krane et al., *Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation*, 53 J. Forensic Sci. 1006 (2008), https://doi.org/10.1111/j.1556-4029.2008.00787.x.

153. *See, e.g.,* State v. Murray, 174 P.3d 407, 417–18 (Kan. 2008) (inconclusive results were presented as consistent with the defendant's blood).

154. For example, the two samples would match if someone framed the defendant by putting a sample of defendant's DNA at the crime scene or in the container of DNA thought to have come from the crime scene. Or, the defendant's DNA might have made its way to the crime-scene sample inadvertently, as is thought to have happened when a homeless man's DNA was found on the fingernails of a multi-millionaire found dead in his home after a robbery. After a public defender established that the defendant was in a hospital when the murder occurred, police and prosecutors concluded that his DNA probably had been transferred by paramedics who treated the defendant three hours before they applied a pulse oximeter to the victim's finger. Katie Worth, Framed for Murder by His Own DNA, Frontline, Apr. 19, 2018, https://perma.cc/C5ST-QWNS.

possibility is laboratory error—the genotypes are not actually the same even though the expert thinks that they are. This situation could arise from mistakes in labeling or handling samples or from cross-contamination of the samples. Another alternative is that the genotypes are truly identical but the forensic sample came from another individual. In general, the true source might be a close relative of the defendant[155] or an unrelated person who just happens to have the same profile as the defendant. The former hypothesis we shall refer to as kinship, and the latter as coincidence. To infer that the defendant is the source of the trace-evidence DNA, one must reject these alternative hypotheses of laboratory error, kinship, and coincidence. Table 1 summarizes the logical possibilities.[156]

Table 1. Hypotheses that Might Explain a Match Between Defendant's DNA and DNA at a Crime Scene

IDENTITY	
Defendant is source	Same genotype, defendant's DNA at crime scene
NONIDENTITY	
Laboratory error	Different genotypes mistakenly found to be the same
Kinship	Same genotype, relative's DNA at crime scene
Coincidence	Same genotype, unrelated individual's DNA at crime scene

If laboratory error, kinship, and coincidence are rejected as implausible, then only the hypothesis of identity remains. We turn, then, to the considerations that affect the chances of a match when the defendant is not the source of the trace evidence.

Can the Match Be Attributed to Laboratory Error?

Traditional legal and scientific procedures can help to assess the possibilities of errors in handling or analyzing the samples. Scrutinizing the chain of custody, examining the laboratory's protocol, verifying that it adhered to that protocol, and conducting confirmatory tests (including testing by the defense) can help show that the profiles really do match. Yet, "[e]rrors happen, even in the best laboratories, and even when the analyst is certain that every precaution against error was taken."[157]

155. A close relative, for these purposes, would be a brother, uncle, nephew, etc. For relationships more distant than second cousins, the probability of a chance match is nearly as small as for persons of the same ethnic subgroup. Bernard Devlin & Kathryn Roeder, *DNA Profiling: Statistics and Population Genetics*, *in* 1 Modern Scientific Evidence: The Law and Science of Expert Testimony § 18–3.1.3, at 724 (David L. Faigman et al. eds., 1997) (section not included in later editions).

156. *Cf.* Newton E. Morton, *The Forensic DNA Endgame*, 37 Jurimetrics J. 477, 480 tbl. 1 (1997).

157. NRC I, *supra* note 6, at 89; *see also* Ate Kloosterman et al., *Error Rates in Forensic DNA Analysis: Definition, Numbers, Impact and Communication*, 12 Forensic Sci. Int'l: Genetics 775 (2014), https://doi.org/10.1016/j.fsigen.2014.04.014.

Some commentary proposes using the proportion of false positives that the particular examiner or laboratory has experienced in blind proficiency tests, or the rate of false positives on proficiency tests averaged across all laboratories, to estimate the probability of a false inclusion in the case at bar.[158] Other commentators stress that there are too few proficiency tests to estimate averages accurately and question the application of an historical industry-wide error rate to a particular laboratory at a later time.[159] Proponents of using averages from proficiency tests reply that this information at least can be used to provide an upper bound on the false-positive laboratory-error probability.[160]

Could a Close Relative Be the Source?

With enough loci to test, all individuals except some identical twins should be distinguishable. With the STR-CE technology in general use and with especially small quantities of DNA, however, this ideal is not always attainable. A thorough investigation might extend to all known relatives, but this is not always feasible, and there is always the chance that some unknown relatives are in the suspect population. Formulas are available for computing the probability that any person with a specified degree of kinship to the defendant also possesses the incriminating genotype.[161] For example, the probability that an untested brother (or sister) would match at four loci (with alleles that each occur in 10% of the population) is about

158. *E.g.,* Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 Law, Probability & Risk 89 (2013), https://doi.org/10.1093/lpr/mgs013 (proposing blind performance tests specifically designed to estimate error rates rather than proficiency tests for quality assurance); Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence at Trial*, 34 Jurimetrics J. 21, 37–38 (1993); NRC I, *supra* note 6, at 94 ("proficiency tests provide a measure of the false-positive and false-negative rates of a laboratory").

159. NRC II, *supra* note 7, at 85–87; Devlin & Roeder, *supra* note 155, § 18–5.3, at 744–45. Such arguments have not persuaded the proponents of estimating the probability of error from industry-wide proficiency testing. *E.g.*, Jonathan J. Koehler, *Why DNA Likelihood Ratios Should Account for Error (Even When a National Research Council Report Says They Should Not)*, 37 Jurimetrics J. 425 (1997).

160. *E.g.*, Jonathan J. Koehler, *DNA Matches and Statistics: Important Questions, Surprising Answers*, 76 Judicature 222, 228 (1993); Richard Lempert, *After the DNA Wars: Skirmishing with NRC II*, 37 Jurimetrics J. 439, 447–48, 453 (1997). For example, if there were no errors in 100 tests, a 95% confidence interval would include the possibility that the error rate could be almost as high as 3%. *See* NRC II, *supra* note 7, at 86 n.1. For an explanation of confidence intervals and the "zero-numerator problem," see Kaye & Stern, *supra* note 12; *see also supra* section titled "What Are DNA Polymorphisms and How Are They Detected?"

161. John S. Buckleton et al., *Relatedness*, *in* Forensic DNA Evidence Interpretation 119 (John S. Buckleton et al. eds., 2d ed. 2016); Bruce S. Weir, *Kinship*, *in* Handbook of Forensic Statistics 265, 268–71 (David Banks et al. eds., 2021) (but also noting the uncertainties or complications in applying the formulas).

1/380;[162] the probability that an aunt (or uncle) would match is about 1/100,000.[163] With more independent loci, the probabilities will plummet.

Could an Unrelated Person Be the Source?

Another rival hypothesis is coincidence: The defendant is not the source of the crime-scene DNA but happens to have the same genotype as an unrelated individual who is the true source. In principle, this hypothesis could be addressed by genotyping everyone in the suspect population. But that is rarely feasible.[164] Instead, an estimate of how frequently the incriminating genotype occurs in broad racial or ethnic populations is provided.[165] The first step is to estimate the frequencies of individual alleles from samples of various populations (often referred to as reference or population databases).[166] Typically, convenience samples (often from blood banks

162. For a case with conflicting calculations of the probability of an untested brother having a matching genotype, see *McDaniel v. Brown*, 558 U.S. 120 (2010) (per curiam). The correct computation is given in David H. Kaye, *"False, but Highly Persuasive": How Wrong Were the Probability Estimates in* McDaniel v. Brown*?,* 108 Mich. L. Rev. First Impressions 1 (2009), https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1059&context=fac_works; and David H. Kaye, *The Interpretation of DNA Evidence: A Case Study in Probabilities*, *in* Science Policy Decision-Making Educational Modules (Nat'l Academies of Sci., Engineering, and Med. Comm. on Preparing the Next Generation of Policy Makers for Science-Based Decisions ed. 2016), https://perma.cc/D9RM-DYDK [hereinafter Kaye, *Case Study*].

163. These figures follow from the equations in NRC II, *supra* note 7, at 113. The large discrepancy between two siblings on the one hand, and an uncle and nephew on the other, reflects the fact that the siblings have far more shared ancestry. All their genotypes are inherited through the same two parents. In contrast, a nephew and an uncle inherit from two unrelated mothers, and so will have few maternal alleles in common. As for paternal alleles, the nephew inherits not from his uncle, but from his uncle's brother, who shares by descent only about one-half of his alleles with the uncle.

164. The suspect population normally defies enumeration, and in the typical crime where DNA evidence is found, the population of possible perpetrators is so huge that even if all of its members could be listed, they could not all be tested. As the cost of DNA profiling drops, however, it will become technically and economically feasible to have a comprehensive, population-wide DNA database that could be used to produce a list of nearly everyone whose DNA profile is consistent with the trace-evidence DNA. Whether such a system would be constitutionally and politically acceptable is another question. *See* David H. Kaye & Michael S. Smith, *DNA Identification Databases: Legality, Legitimacy, and the Case for Population-Wide Coverage*, 2003 Wis. L. Rev. 413. But the introduction of investigative genetic genealogy has renewed interest in the concept. J.W. Hazel et al., *Is It Time for a Universal Genetic Forensic Database?*, 362 Science 898 (2018).

165. The underlying premise is that, considering the biology of DNA markers, each possible forensic DNA genotype occurs just as often among criminal offenders as anybody else. So if we know the profile frequency in a large, relevant population group, we can use it for the unobserved set of conceivable suspects.

166. In the formative years of forensic DNA testing, defendants frequently contended that forensic databases were too small to give accurate estimates, but this argument generally proved unpersuasive. *E.g.*, United States v. Shea, 957 F. Supp. 331, 341–43 (D.N.H. 1997); State v. Copeland, 922 P.2d 1304, 1321 (Wash. 1996). To the extent that the databases are comparable to random

or paternity cases, with the "race" of the donor being self-declared or assessed by the person collecting the sample) are used to construct the databases.[167] Second, the estimated allele frequencies at each locus are combined into single-locus frequencies. The simplest computation posits an infinite population of individuals who choose their mates and reproduce independently of the forensic-identification alleles they possess.[168] Then the expected population proportion of a pair of alleles (a single-locus genotype) is essentially the product of allele frequencies.[169] Finally, the single-locus frequencies are multiplied together.[170]

Because the frequencies vary across census groups (White, Black, Hispanic, Asian, and Native American), it became common to present separate estimates for these groups, at least in cases in which the race of the source of the trace evidence was unknown. For example, if a woman was abducted from a rest stop on an interstate highway and sexually assaulted, the suspect population could include people of all census categories.[171] Random-match probabilities for all the major

samples, confidence intervals can be used to indicate the uncertainty related to sample size. Unfortunately, the meaning of a confidence interval is subtle, and the estimate commonly is misconstrued. *See* Kaye & Stern, *supra* note 12.

167. A few experts have testified that no meaningful conclusions can be drawn in the absence of random sampling. *E.g.,* People v. Soto, 88 Cal. Rptr. 2d 34 (1999); State v. Anderson, 881 P.2d 29, 39 (N.M. 1994). The 1996 NRC report suggests that for the purpose of estimating allele frequencies, convenience sampling should give results comparable to random sampling, and it discusses procedures for estimating the random sampling error. NRC II, *supra* note 7, at 126–27, 146–48, 186. The courts generally have rejected the argument that random samples are essential to valid or generally accepted random-match probabilities. *See* section titled "A Brief History of DNA Evidence" above.

168. Of course, people do not choose their mates by a lottery. "Random mating" simply indicates that the choices are uncorrelated with the specific alleles that make up the genotypes in question. A further condition of the random-mating model is that there is no systematic relationship between the particular alleles that an offspring receives and the chance that the child will transmit an allele to the next generation.

169. If the single-locus genotype is homozygous (for example, the allele from each parent is A_1), then expected proportion is just $p_1 \times p_1$, or p_1^2, where p_1 is the proportion of all alleles in the population that are type A_1. If the pair of alleles at a locus are distinct (call them A_1 and A_2), then the single-locus heterozygous genotype proportion is $2p_1p_2$. For example, suppose that 10% of the sperm in the gene pool of the population carry allele 1, and 50% carry allele 2. Similarly, 10% of the eggs carry A_1, and 50% carry A_2. (Other sperm and eggs carry other types.) With random mating, we expect $10\% \times 10\% = 1\%$ of all the fertilized eggs to be A_1A_1, and another $50\% \times 50\% = 25\%$ to be A_2A_2. These constitute two distinct homozygote profiles. Likewise, we expect $10\% \times 50\% = 5\%$ of the fertilized eggs to be A_1A_2 and another $50\% \times 10\% = 5\%$ to be A_2A_1. These two configurations produce indistinguishable profiles—a peak, band, or dot for A_1 and another mark for A_2. So the expected proportion of heterozygotes is $5\% + 5\% = 10\%$.

These proportions are known as Hardy-Weinberg proportions. Even if two populations with distinct allele frequencies are thrown together, within the limits of chance variation, random mating produces Hardy-Weinberg equilibrium in a single generation.

170. When the various single-locus genotypes are statistically independent of one another, the population is said to be in linkage equilibrium.

171. United States v. Jakobetz, 955 F.2d 786 (2d Cir. 1992).

groups enable the jury to see how much the estimates vary for different parts of the population even if they cannot determine which of these parts is most salient.

In cases in which the suspect population seems limited to a single major population group—for example, where the victim is sure that the assailant looked Hispanic and spoke with a Spanish accent—only the frequency estimated for that group might be presented.[172] But what if the victim is mistaken or the perpetrator is of mixed ancestry? If the wrong racial or ethnic database were used to estimate a profile frequency, it would yield too small (or too large) a random-match probability. Thus, even in cases in which the evidence points to a subpopulation, there is an argument for giving the estimates for several groups.[173]

A concern often raised with racial classifications involving DNA evidence is that references to race and DNA genotypes of any kind will "reify" the idea that racial categories are fundamentally biological rather than socially constructed.[174] There are ways to avoid the census-type labels. Allele frequency data can be collected from regions across the globe, and the spectrum of resulting profile frequency estimates could be presented in every case.[175] Alternatively, only the highest frequency estimate across this spectrum could be selected to supply a "conservative" estimate (that is, an overestimate). One might even create an unrealistic population by picking, for each and every allele, the frequency that is the largest in any of the groups and combining those maximum frequencies.[176] Ideas like these have been proposed,[177] and

172. *Cf.* People v. Pizarro, 12 Cal. Rptr. 2d 436, 441 (Ct. App. 1992), *after remand*, 3 Cal. Rptr. 3d 21 (Ct. App. 2003), *review denied* (Oct. 15, 2003).

173. David H. Kaye, *The Role of Race in DNA Evidence: What Experts Say, What California Courts Allow*, 37 Sw. U. L. Rev. 303 (2008).

174. *E.g.*, Troy Duster, *Race and Reification in Science*, 307 Science 1050 (2005); section titled "Biogeographic Ancestry Testing" below.

175. The FBI's STR allele-frequency database has samples from "African Americans, Caucasians, Southeast Hispanics, Southwestern Hispanics, Bahamians, Jamaicans, Trinidadians, Apaches, Navajos, Chamorros and Filipinos." Tamyra R. Moretti et al., *Population Data on the Expanded CODIS Core STR Loci for Eleven Populations of Significance for Forensic DNA Analyses in the United States*, 25 Forensic Sci. Int'l: Genetics 175, 175 (2016), https://doi.org/10.1016/j.fsigen.2016.07.022.

176. This is a rough statement of one of the controversial "ceiling" methods "strongly recommend[ed]" in NRC I, *supra* note 6, at 82–85.

177. An advisory commission appointed by the Department of Justice noted that a more refined population-genetics model (mentioned below) could be used to treat the entire United States as a single, structured population, obviating the reporting of estimates by more specific groups. Nat'l Comm'n on the Future of DNA Evidence, The Future of DNA Evidence: Predictions of the Research and Development Working Group 5 (2000); *see also* Robert F. Oldt & Sreetharan Kanthaswamy, *Expanded CODIS STR Allele Frequencies—Evidence for the Irrelevance of Race-based DNA Databases*, 42 Legal Med. 101642 (2020), https://doi.org/10.1016/j.legalmed.2019.101642 (questioning the value of racial labels when typing at 20 STR loci is successful). Another procedure that avoids racial categories is simply to give the (larger) probability of a random match to a brother or sister at the loci used in the testing. Because full siblings always are more closely related than are

some were used in cases.[178] However, the dominant reason given for looking to more narrowly defined populations was a perception that the model of randomly mating major racial populations was just too simple. Perhaps linguistic, religious, or other subgroups within the broader populations mate almost exclusively among themselves. If these subpopulations happen to have varying allele frequencies, then the estimates for the DNA-genotype frequencies in the broader groups could understate (or overstate) the true frequencies for the racial groups or be inapposite in a case in which the suspect population is confined to a narrow subgroup.

For a time, the likely magnitude of the effect of this "population structure" on profile-frequency estimates was hotly contested.[179] Today, a quantity denoted as F_{ST} or θ (theta)[180] often is used in formulas designed to account for population structure;[181] however, there is a continuing discussion of what value to use for θ,[182] and some researchers have developed further refinements.[183]

Frequencies, Probabilities, and Prejudice

Up to this point, we have described the estimates for genotype frequencies (often presented as random-match probabilities) that commonly are quoted in conjunction with the finding that the trace-evidence sample contains DNA of the same

randomly drawn pairs from even a narrowly defined subpopulation, this "Sib Method . . . provides a rough upper limit for the actual match probability." Nat'l Comm'n, *supra*, at 5.

178. *E.g.*, Brim v. State, 695 So.2d 268 (Fla. 1997).

179. Jay D. Aronson, Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling (2007); Kaye, *supra* note 1, at 121–26. By the mid-1990s, the population-structure objection to admitting random-match probabilities had lost its punch. *See, e.g.,* Kaye, *supra* note 1; *supra* section titled "A Brief History of DNA Evidence."

180. *See* David Balding & Richard A. Nichols, *DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands*, 64 Forensic Sci. Int'l 125 (1994); Gill et al., *supra* note 42, at 29–38.

181. The recommendations of the 1996 NAS committee for computing random-match probabilities with these formulas for broad populations and particular subpopulations are summarized in the second edition of this guide. *See* Reference Manual on Scientific Evidence (2d ed. 2000).

182. *See* John Buckleton et al., *Population-specific* FST *Values for Forensic STR Markers: A Worldwide Survey*, 23 Forensic Sci. Int'l: Genetics 91 (2016); Bruce S. Weir, *DNA Frequencies and Probabilities*, *in* Handbook of Forensic Statistics 251, 256–57 (David Banks et al. eds., 2021); Weir, *supra* note 161, at 266–68. The U.K. Forensic Science Regulator recommends the value of 0.03 as "sufficiently conservative that it is almost certainly favourable to defendants even allowing for alternative contributors to come from very different ethnic populations." Forensic Science Regulator (U.K.), Guidance: Allele Frequency Databases and Reporting Guidance for the DNA (Short Tandem Repeat) Profiling, FSR-G-213 Issue 2, § 14.1.4 (2020); however, "in unusual cases involving small and isolated populations that may be highly differentiated from available databases," 0.05 could be used. *Id.* § 14.1.1.

183. *See* Mark A. Jobling, *Forensic Genetics Through the Lens of Lewontin: Population Structure, Ancestry and Race*, 377 Phil. Transactions Royal Soc'y B 20200422 (2022) (citing three papers).

type as the defendant's. Assuming that the statistical methods meet *Daubert*'s demand for scientific validity and reliability, and thus satisfy Federal Rule of Evidence 702 (or, in some states, *Frye*'s requirement of general acceptance in the scientific community), a further issue can arise under Rule 403: To what extent will the presentation assist the jury in understanding the meaning of a match so that the jury can give the evidence the weight that it deserves? This question involves psychology and law, and we summarize the arguments about probative value and prejudice that have been made in litigation and in the legal and scientific literature. How the legal issue of the admissibility of any particular statistic generally should be resolved under the balancing standard of Rule 403 may turn not only on the general features of the evidence described here, but on the context and circumstances of particular cases.

Are Frequencies or Probabilities Prejudicial Because They Are So Small?

The most common form of expert testimony about matching DNA involves an explanation of how the laboratory ascertained that the defendant's DNA has the profile of the trace-evidence sample plus an estimate of the profile frequency or random-match probability.[184] It has been suggested, however, that jurors do not understand probabilities in general, and that infinitesimal match probabilities will so bedazzle jurors that they will not appreciate the other evidence in the case or any innocent explanations for the match.[185] Empirical research into this hypothesis does not clearly support the argument that jurors will overweight the probability.[186] The details of how the probability is presented and countered may be important, and remedies short of exclusion are available.[187] The practice in

184. A more flexible alternative that has widespread support among forensic DNA scientists and forensic statisticians is the likelihood ratio or Bayes' factor discussed *infra* section titled "Mixtures."

185. *Cf.* Gov't of the Virgin Islands v. Byers, 941 F. Supp. 513, 527 (D.V.I. 1996) ("Vanishingly small probabilities of a random match may tend to establish guilt in the minds of jurors and are particularly suspect.").

186. Thomas Busey et al., *Validating Strength-of-Support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions*, 67 J. Forensic Sci. 936 (2022) ("whether jurors can understand more complex statistical terms such as likelihood ratios and random match probabilities is an empirical question with no clear answer in the literature"); *see generally* Kristy A. Martire, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists*, *in* Handbook of Forensic Statistics 201 (David Banks et al. eds., 2021); David H. Kaye et al., *Statistics in the Jury Box: Do Jurors Understand Mitochondrial DNA Match Probabilities?*, 4 J. Empirical Legal Stud. 797 (2007).

187. United States v. Graves, 465 F. Supp. 2d 450, 458 (E.D. Pa. 2006) ("with cross-examination, proper explanations, and clarifying jury instructions, the probative value of the sneaker evidence [with RMPs of about 1/3,000] is not substantially outweighed by the danger of unfair prejudice and confusion"); United States v. Santiago, 156 F. Supp. 2d 145, 151 (D.P.R.

the United Kingdom is to cap reported random-match probabilities at "1 in 1 billion."[188]

Thus, although there once was a line of cases that excluded probability testimony in criminal matters, by the mid-1990s, no jurisdiction excluded DNA match probabilities on this basis.[189] The opposite argument—that relatively large random-match probabilities are prejudicial because jurors will not appreciate that they make the DNA match unimpressive—also has been advanced without much success.[190]

Are Frequencies or Probabilities Prejudicial Because They Might Be Transposed?

A related concern is that the jury will misconstrue the random-match probability as the probability that the evidence DNA came from a random individual. The words are almost identical, but the probabilities can be different. The random-match probability is the probability that the suspect's DNA matches the trace-evidence sample calculated on the proposition that the suspect is not the true source of that sample (and is not a close relative). The proposition is a hypothesis about who the source really is; it asserts that the true source is someone other than the suspect. But the laboratory cannot evaluate the probability of this hypothesis on the basis of

2001); NRC II, *supra* note 7, at 197 (suitable cross-examination, defense experts, and jury instructions might reduce the risk of an unwarranted sense of certainty).

188. Forensic Science Regulator, Guidance: Allele Frequency Databases and Reporting Guidance for the DNA (Short Tandem Repeat) Profiling (2014) (FSR-G-213 Issue 1, Recommendation 4, retained in Issue 2, 2020, §§ 8 & 10). In part, this limit reflects the concern that the assumptions of population genetics and statistical models do not hold exactly, reducing the probative value of much smaller numbers. It also is in the spirit of the 2023 amendment to Rule 702 intended "to emphasize that . . . [a] testifying expert's opinion must stay within the bounds of what can be concluded by a reliable application of the expert's basis and methodology." Fed. R. Evid. 702 advisory committee's note to the 2023 amendment. *But see* Tacha Hicks et al., *A Logical Framework for Forensic DNA Interpretation*, 13 Genes 957 (2022), https://doi.org/10.3390/genes13060957 (Noting that when "the Forensic Science Service introduced the ten locus STR system into casework," the one-in-a-billion cap was used as "a temporary expedient" because it was "argued that the extent of investigations into the independence assumptions . . . was insufficient to justify the robustness of LRs in excess of one billion"; yet "to this day, there still seems to be little in the way of large-scale between-locus dependence effects," making the practice for the much larger "number of loci now in routine use . . . a peculiar state of affairs.").

189. *E.g.,* United States v. Chischilly, 30 F.3d 1144 (9th Cir. 1994); State v. Weeks, 891 P.2d 477, 489 (Mont. 1995); *see also* State v. Bauldwin, 811 N.W.2d 267, 288 (Neb. 2012).

190. *E.g.,* United States v. McCluskey, 954 F. Supp. 2d 1224, 1273–75 (D.N.M. 2013); State v. Tucker, 920 N.W.2d 680, 688–89 (Neb. 2018). *But see* United States v. Graves, *supra* note 187, at 459 ("even with appropriate safeguards, the minimal probative value of the umbrella DNA evidence—in which half of the relevant population cannot be excluded as a contributor to the DNA sample—is substantially outweighed by the danger of unfair prejudice and confusion of the issues").

the DNA evidence itself. It can only provide the probability of an event such as the suspect's DNA profile matching the crime-scene DNA profile if the true source is genetically unrelated to the suspect (or related in a specified manner). The occurrence of the match certainly is circumstantial evidence in favor of the hypothesis that the suspect is the source; however, automatically "equating source probability with random match probability" is "faulty reasoning."[191]

Sliding from the probability of the circumstantial evidence to the probability of the circumstances themselves (the hypotheses about the evidence) often is called the prosecutor's fallacy,[192] although instances are hardly confined to prosecutors.[193] Statisticians know it as the fallacy of the transposed conditional.[194] Despite the attention the transposition fallacy has received, no federal court has excluded a random-match probability as unfairly prejudicial simply because the jury might misinterpret it as a probability that the defendant is the

191. McDaniel v. Brown, 558 U.S. 120, 128 (2010) (per curiam).

192. *Id*. In *Brown*, a DNA analyst, at the prompting of the prosecution, suggested that a random-match probability of 1/3,000,000 implied a 0.000033 probability that the defendant was not the source of the DNA found on the victim's clothing. The Court unanimously held that a federal writ of habeas corpus should not have been issued in light of the limitations on federal review of state convictions. The prisoner argued, inter alia, that the transposed conditional probability made the DNA evidence so unreliable that its use deprived him of due process. The Court did not reach this issue, as it had not been raised below. The probabilities associated with the DNA evidence in the case are discussed in detail in Kaye, *Case Study*, *supra* note 162.

193. It also appears in statements from defense counsel, scientists, journalists, and judges. *E.g.,* Williams v. Bauman, 759 F.3d 630, 632 (6th Cir. 2014) ("a 1–in–7.663 quadrillion chance that it came from someone else"); United States v. Ford, 683 F.3d 761, 768 (7th Cir. 2012) ("[t]he probability that the DNA was someone else's would be 7 percent if the comparison were confined to the first location"); *see also* David H. Kaye et al., The New Wigmore, A Treatise on Evidence: Expert Evidence § 14.1.2(a) (2d ed. 2011) (chapter not included in the current edition) (collecting opinions expressing the fallacy); 2 Paul C. Giannelli et al., Scientific Evidence § 18.04[3][b][4], at 1893 n.384 ("nearly every appellate decision in the U.S. on the sufficiency of a DNA cold hit has fallen victim to the fallacy").

There is an opposing "defendant's fallacy" of dismissing or undervaluing matches because other matches are to be expected in unrealistically large populations of potential suspects. For example, defense counsel might argue that (1) with a random-match probability of one in a million, we would expect to find three or four unrelated people with the requisite genotypes in a major metropolitan area with a population of 3.6 million; (2) the defendant just happens to be one of these three or four, which means that the chances are at least 2 out of 3 that someone unrelated to the defendant is the source; so (3) the DNA evidence does nothing to incriminate the defendant. The problem with this argument is that in a case involving both DNA and non-DNA evidence against the defendant, it is unrealistic to assume that there are 3.6 million equally likely suspects. When juries are confronted with both fallacies, the defendant's fallacy seems to dominate. NRC II, *supra* note 7, at 198; *cf.* Jonathan J. Koehler, *The Psychology of Numbers in the Courtroom: How to Make DNA-Match Statistics Seem Impressive or Insufficient*, 74 S. Cal. L. Rev. 1275 (2001) (discussing ways of framing the evidence that make it more or less persuasive).

194. *See* Kaye & Stern, *supra* note 12, app'x section C (describing the fallacy and the relationship between the conditional probabilities); Ian W. Evett, *Avoiding the Transposed Conditional*, 35 Sci. & Just. 127 (1995).

source of the forensic DNA.[195] Courts, however, have noted the need to have the concept "properly explained,"[196] and prosecutorial or expert misrepresentations of the random-match probabilities for DNA and other trace evidence have produced reversals or contributed to the setting aside of verdicts.[197]

Are Random-Match Probabilities That Are Smaller Than False-Positive Error Probabilities Irrelevant or Prejudicial?

Some scientists and lawyers have maintained that match probabilities are logically irrelevant when they are far smaller than the probability of a frame-up, a blunder in labeling samples, cross-contamination, or other events that would

195. *See, e.g.,* United States v. McCluskey, 954 F. Supp. 2d 1224, 1291 (D.N.M. 2013); United States v. Morrow, 374 F. Supp. 2d 51, 66 (D.D.C. 2005) ("careful oversight by the district court and proper explanation can easily thwart this issue").

196. United States v. Shea, 957 F. Supp. 331, 345 (D.N.H. 1997); *see also* United States v. Chischilly, 30 F.3d 1144, 1158 (9th Cir. 1994) (stating that the government must be "careful to frame the DNA profiling statistics presented at trial as the probability of a random match, not the probability of the defendant's innocence that is the crux of the prosecutor's fallacy"). The 1996 NRC committee suggested that "if the initial presentation of the probability figure, cross-examination, and opposing testimony all fail to clarify the point, the judge can counter [the fallacy] by appropriate instructions to the jurors that minimize the possibility of cognitive errors." NRC II, *supra* note 7, at 198 (footnote omitted). The committee formulated the following instruction to define the random-match probability:

> In evaluating the expert testimony on the DNA evidence, you were presented with a number indicating the probability that another individual drawn at random from the [specify] population would coincidentally have the same DNA profile as the [bloodstain, semen stain, etc.]. That number, which assumes that no sample mishandling or laboratory error occurred, indicates how distinctive the DNA profile is. It does not by itself tell you the probability that the defendant is innocent.

Id. at 198 n.93. An alternative adopted in England is to confine the prosecution to stating a frequency rather than a probability. *See* Kaye et al., *supra* note 193, § 14.1.2(b); *cf.* D.H. Kaye, *The Admissibility of "Probability Evidence" in Criminal Trials—Part II*, 27 Jurimetrics J. 160, 168 (1987) (similar proposal).

197. *E.g.,* United States v. Massey, 594 F.2d 676, 681 (8th Cir. 1979) (explaining that in closing argument about hair evidence, "the prosecutor 'confuse[d] the probability of concurrence of the identifying marks with the probability of mistaken identification'") (alteration in original); *cf.* Whack v. State, 73 A.3d 186, 198 (Md. 2013) (reversing defendant's conviction because of the prosecution's closing argument about two probabilities and admonishing that "counsel have a responsibility to take extra care in describing DNA evidence, particularly when it comes to statistical probabilities"); State v. Phillips, 844 S.E.2d 651, 658 (S.C. 2020) (reversing partly as a result of inadequacies and errors in the presentation of random-match probabilities for "touch" DNA evidence and noting that "even when the concepts of touch DNA, non-exclusion DNA, and random match probability are completely and accurately presented to a jury, there is significant potential the testimony will be confusing and misleading").

yield a false positive.[198] The argument is that the jury should concern itself only with the chance that the forensic sample is reported to match the defendant's profile even though the defendant is not the source. Match probabilities address only one path to such a false-positive report—a coincidence in correctly ascertained profiles. They do not consider the probability of a false-positive report because of fraud or an error in the collection, handling, or analysis of the DNA samples. Unless those probabilities are essentially zero, the match probability understates the chance of a reported match arising when DNA from a nonsource is compared to DNA from the source of the trace-evidence sample.

This mathematical observation has led to arguments that because probability of these other possible explanations for a match are larger than the very small random-match probabilities for most STR profiles, the latter probabilities are irrelevant. Commentators have crafted theoretical, doctrinal, and practical rejoinders to this claim.[199] The essence of the counterargument is that it is logical to give jurors information about kinship or random-match probabilities because, even if these numbers do not give the whole picture, they address pertinent hypotheses about the true source of the trace evidence.

It also has been argued that even if very small match probabilities are logically relevant, they are unfairly prejudicial in that they will cause jurors to neglect the probability of a match arising because of a false-positive laboratory error.[200] A court that shares this concern might require the expert who presents a random-match probability also to report a probability that the laboratory is mistaken about

198. *E.g.,* Jonathan J. Koehler et al., *The Random Match Probability in DNA Evidence: Irrelevant and Prejudicial?*, 35 Jurimetrics J. 201 (1995); Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 Science 1745, 1749 (1991), https://doi.org/10.1126/science.1845040 ("[p]robability estimates like 1 in 738,000,000,000 . . . are terribly misleading because the rate of laboratory error is not taken into account").

199. *See* Kaye et al., *supra* note 193, § 14.1.1 (discussing the issue).

200. Some commentators believe that this prejudice is so likely and so serious that "jurors ordinarily should receive *only* the laboratory's false positive rate. . . ." Richard Lempert, *Some Caveats Concerning DNA as Criminal Identification Evidence: With Thanks to the Reverend Bayes*, 13 Cardozo L. Rev. 303, 325 (1991) (emphasis added). The 1996 NRC committee was skeptical of this view, especially when the defendant has had a meaningful opportunity to retest the DNA at a laboratory of his or her choice, and it suggested that judicial instructions can be crafted to avoid this form of prejudice. NRC II, *supra* note 7, at 199. Pertinent psychological research includes Dale A. Nance & Scott B. Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random Match Probability*, 34 J. Legal Stud. 395 (2005); Dale A. Nance & Scott B. Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability*, 42 Jurimetrics J. 403 (2002); Jason Schklar & Shari Seidman Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 Law & Hum. Behav. 159, 179 (1999), https://doi.org/10.1023/A:1022368801333 (concluding that separate figures for laboratory error and a random match to a correctly ascertained profile are desirable in that "[j]urors . . . may need to know the disaggregated elements that influence the aggregated estimate as well as how they were combined in order to evaluate the DNA test results in the context of their background beliefs and the other evidence introduced at trial").

the profiles. Of course, for reasons given in the section titled "How Are Samples Created and Handled?," above, some experts would deny that they can provide a meaningful statistic for the case at hand, but they could report the results of proficiency tests and leave it to the jury to use this figure as best it can in considering whether a false-positive error has occurred.[201] In any event, the courts have been unreceptive to efforts to replace random-match probabilities with a blended figure that incorporates the risk of a false-positive error[202] or to exclude random-match probabilities that are not accompanied by a separate false-positive error probability.[203]

"Rarity," Source, and Uniqueness Testimony

Having surveyed the issues related to the value and dangers of probabilities and statistics for DNA evidence, we turn to a related issue that can arise under Rules 702 and 403: Should an expert be permitted to offer a nonnumerical judgment about the DNA profiles? Many courts have held that a DNA match is inadmissible unless the expert attaches a scientifically valid number to the match. Indeed, some opinions state that this requirement flows from the nature of science

201. *Cf.* Williams v. State, 679 A.2d 1106, 1120 (Md. 1996) (reversing because the trial court restricted cross-examination about the results of proficiency tests involving other DNA analysts at the same laboratory). *But see* United States v. Shea, 957 F. Supp. 331, 344 n.42 (D.N.H. 1997) ("The parties assume that error rate information is admissible at trial. This assumption may well be incorrect. Even though a laboratory or industry error rate may be logically relevant, a strong argument can be made that such evidence is barred by Fed. R. Evid. 404 because it is inadmissible propensity evidence.").

202. United States v. McCluskey, 954 F. Supp. 2d 1224, 1270–73 (D.N.M. 2013); United States v. Ewell, 252 F. Supp. 2d 104, 113–14 (D.N.J. 2003); United States v. Shea, 957 F. Supp. 331, 334–45 (D.N.H. 1997); People v. Reeves, 109 Cal. Rptr. 2d 728, 753 (Ct. App. 2001); State v. Tester, 968 A.2d 895 (Vt. 2009); *cf.* Armstead v. State, 673 A.2d 221, 245 (Md. 1996) (the failure to combine a random-match probability with an error rate on proficiency tests that was many orders of magnitude greater (and that was placed before the jury) did not deprive the defendant of due process). Formulas for incorporating error into the likelihood are given in Tacha Hicks et al., *A Framework for Interpreting Evidence*, *in* Forensic DNA Evidence Interpretation 37, 73 (John Buckleton et al. eds., 2d ed. 2016), https://doi.org/10.1201/b19680-3.

203. *McCluskey*, 954 F. Supp. 2d at 1270–73; United States v. Trala, 162 F. Supp. 2d 336, 350–51 (D. Del. 2001); United States v. Lowe, 954 F. Supp. 401, 415–16 (D. Mass. 1997), *aff'd*, 145 F.3d 45 (1st Cir. 1998) (a "theoretical" error rate need not be presented when quality-assurance standards have been followed and defendant had the opportunity to retest the sample); Roberts v. United States, 916 A.2d 922, 930–31 (D.C. 2007); Roberson v. State, 16 S.W.3d 156, 168 (Tex. Crim. App. 2000) (error rate not needed when laboratory was accredited and underwent blind proficiency testing); *Tester*, 968 A.2d 895 (finding when the laboratory chemist stated that "[t]here is no error rate to report" because the number of proficiency trials was insufficient, the random-match probability was admissible and preferable to presenting the finding of a match with no accompanying statistic).

itself.[204] However, this view has been challenged,[205] and not all courts agree that an expert must explain the power of a DNA match in purely numerical terms.[206]

Instead of presenting numerical frequencies or match probabilities, a scientist could characterize a multilocus STR profile as "rare," "extremely rare," or the like.[207] A few opinions suggest that a purely qualitative presentation would be admissible,[208] but with the availability of statistically sound estimates of frequencies or conditional probabilities, such testimony is itself quite rare. When numerical estimates are possible, it is not clear what the expert's verbal tag accomplishes.

The most extreme case of a purely verbal description of the infrequency of a profile occurs when that profile can be said to be unique. Of course, the uniqueness of any object, from a snowflake to a fingerprint, in a population that cannot be enumerated, never can be proved directly. As with all sample evidence, one must generalize from the sample to the entire population. There is always some probability that a census would prove the generalization to be false. Almost three decades ago, the second National Research Council (NRC) committee therefore wrote:

> [T]here is no "bright-line" standard in law or science that can pick out exactly how small the probability of the existence of a given profile in more than one member of a population must be before assertions of uniqueness are justified. . . . There might already be cases in which it is defensible for an expert to assert that, assuming that there has been no sample mishandling or laboratory error, the profile's probable uniqueness means that the two DNA samples come from the same person.[209]

204. *E.g.*, State v. Cauthron, 846 P.2d 502 (Wash. 1993).

205. *See, e.g.,* State v. Hummert, 933 P.2d 1187, 1191 (1997) ("We believe this view to be seriously incorrect."); Commonwealth v. Crews, 640 A.2d 395, 402 (Pa. 1994) ("[T]he factual evidence of the physical testing of the DNA samples and the matching alleles, even without statistical conclusions, tended to make appellant's presence more likely than it would have been without the evidence, and was therefore relevant."). The National Research Council committee wrote that science only demands "underlying data that permit some reasonable estimate of how rare the matching characteristics actually are," and "[o]nce science has established that a methodology has some individualizing power, the legal system must determine whether and how best to import that technology into the trial process." NRC II, *supra* note 7, at 192.

206. *E.g.*, Rodriguez v. State, 273 P.3d 845, 851 (Nev. 2012); People v. Her, 157 Cal. Rptr. 3d 40 (Cal. Ct. App. 2013). For discussion of pure "defendant-not-excluded" testimony, see *United States v. Morrow*, 374 F. Supp. 2d 51 (D.D.C. 2005); Kaye et al., *supra* note 193, § 15.4.

207. Banks v. State, 219 So.3d 19, 28 (Fla. 2017) ("by the time you get to 13 [STR loci] it becomes extremely rare, extremely"); *cf.* State v. Hummert, 933 P.2d 1187, 1193 (Ariz. 1997) (testimony that "of their own experience, they believed such a [three-locus RFLP-VNTR] random match would be very uncommon" was admissible).

208. State v. Bloom, 516 N.W.2d 159, 166–67 (Minn. 1994) ("Since it may be pointless to expect ever to reach a consensus on how to estimate, with any degree of precision, the probability of a random match and that, given the great difficulty in educating the jury as to precisely what that figure means and does not mean, it might make sense to simply try to arrive at a fair way of explaining the significance of the match in a verbal, qualitative, nonquantitative, nonstatistical way.").

209. NRC II, *supra* note 7, at 194.

Before concluding that a DNA profile is unique in a given population, however, a careful expert also should consider not only the random-match probability (which pertains to unrelated individuals) but also the chance of a match to a close relative. Indeed, the possible existence of an unknown, identical twin also means that a scientist never can be absolutely certain that crime-scene evidence could have come from only the defendant.

Courts have accepted or approved of expert assertions of uniqueness or of individual-source identification.[210] For these assertions to be justified, a large number of sufficiently polymorphic loci must have been tested, making the probabilities of matches to both relatives and unrelated individuals so tiny that the probability of finding another person who could be the source within the relevant population is negligible.[211] But deciding what probability is negligible is a personal or policy judgment rather than an objective fact, and trusting the probability models at the extremes where they cannot be experimentally tested

210. *E.g.,* United States v. Carney, No. 1:20-cr-121, 2022 WL 1155902, at *1 (S.D. Ohio Apr. 19, 2022) ("in the absence of an identical twin, Furious Carney is the source of the major DNA profile"); United States v. Davis, 602 F. Supp. 2d 658 (D. Md. 2009) ("the random match probability figures . . . are sufficiently low so that the profile can be considered unique"); People v. Cordova, 358 P.3d 518, 538 (2015) (because "the odds were astronomical," an "opinion that defendant was the source of the evidentiary samples to a reasonable scientific certainty was reasonable"); State v. Hauge, 79 P.3d 131 (Haw. 2003) (uniqueness); Young v. State, 879 A.2d 44, 46 (Md. 2005) (holding that "when a DNA method analyzes genetic markers at sufficient locations to arrive at an infinitesimal random match probability, expert opinion testimony of a match and of the source of the DNA evidence is admissible"; hence, it was permissible to introduce a report providing no statistics but stating that "(in the absence of an identical twin), Anthony Young (K1) is the source of the DNA obtained from the sperm fraction of the Anal Swab (R1)"); State v. Buckner, 941 P.2d 667, 668 (Wash. 1997) (in light of 1996 NRC Report, "we now conclude there should be no bar to an expert giving his or her expert opinion that, based upon an exceedingly small probability of a defendant's DNA profile matching that of another in a random human population, the profile is unique").

211. Three distinct probabilities arise in speaking of the uniqueness of DNA profiles. First, there is the probability of a match to a single, randomly selected individual in the population. This is the random-match probability. Second, there is the probability that the particular profile is unique. This probability involves pairing the profile with every member of the population. Third, there is the probability that all pairs of all profiles are unique. The first probability is larger than the second, which is many times larger than the third. Uniqueness or source testimony need only establish that *the one* DNA profile in the trace evidence is unique—and not that *all* DNA profiles are unique. Thus, it is the second probability, properly computed, that must be quite small to warrant the conclusion that no one but the defendant (and any identical twins) could be the source of the crime-scene DNA. *See* David H. Kaye, *Identification, Individuality, and Uniqueness: What's the Difference?*, 8 Law, Probability & Risk 85 (2009), https://doi.org/10.1093/lpr/mgp018.

Formulas for estimating all these probabilities are given in NRC II, *supra* note 7, but DNA analysts and judges sometimes infer uniqueness on the basis of incorrect intuitions about the size of the random-match probability. *See* David J. Balding, Weight-of-Evidence for Forensic DNA Profiles 148 (2005) (describing "the uniqueness fallacy"); *cf.* State v. Lee, 976 So. 2d 109, 117 (La. 2008) (incorrect but harmless miscalculation).

makes some scientists wary of claiming uniqueness and making source attributions on the basis of DNA evidence.[212]

Special Issues in Human DNA Testing

Y Chromosomes

Genetic Principles

To understand how Y chromosomes are used in forensic-DNA science, we need to recall a few principles of the genetics of sexual reproduction mentioned in the section above titled "Sexual Reproduction and the Genome." A basic point is that a male child receives an X chromosome from the genetic mother and a Y from the genetic father,[213] whereas female children receive two X chromosomes, one from each parent.[214] Thus, genetic males are type XY and genetic females are XX.[215]

A second fundamental point is that every sex cell (an egg or a sperm) has only one copy of each parental chromosome, selected at random from the pairs carried by each parent. Hence, about half the sperm cells in a genetic male have a Y chromosome, whereas the male's bodily (somatic) cells all have the full XY pair of chromosomes.

Third—and this is crucial in forensic identification with Y chromosomes—when a sperm cell forms, there is limited recombination (exchanges of segments) between X and Y chromosomes. Along most of its length, the Y chromosome stays intact, and only occasional intergenerational mutations produce diversity among Y chromosomes within a population. In other words, the Y loci used in forensic identification are linked rather than independent. They are inherited as a single block—a haplotype—from father to son, and all the men in the same paternal line (up to the last mutation giving rise to a new line in the family tree) would match the trace-evidence sample's Y haplotype. At the same time, men from a different paternal lineage (one with no recent male ancestor in common) might have developed the same haplotype (depending on the history of mutations

212. John S. Buckleton et al., *Single Source Samples*, *in* Forensic DNA Evidence Interpretation 203, 207–15 (John S. Buckleton et al. eds., 2d ed. 2016).

213. All told, the Y chromosome represents about 2% of the total human genome and is much smaller than its X counterpart.

214. This pattern is a consequence of the fact that certain genes in the Y chromosome result in development as a male rather as than a female.

215. Sometimes, progeny are born with fewer or more than the usual two sex chromosomes. Jeannie Visootsak & John M. Graham, Jr., *Klinefelter Syndrome and Other Sex Chromosomal Aneuploidies*, 1 Orphanet J. Rare Diseases 42 (2006), https://doi.org/10.1186/1750-1172-1-42; Brittany Sood & Roselyn W. Clemente Fuentes, *Jacobs Syndrome*, StatPearls (2022), https://perma.cc/62WT-W4S3; *supra* note 22.

giving rise to each line). If so, men from both lines would match a trace-evidence sample with the shared haplotype.

Like the other 22 paired chromosomes (the autosomes), the Y chromosome contains not only genes but also STRs, SNPs, and other parts that vary from one person (or group of people) to the next. A particular haplotype is thus some combination of the variants at some of these loci.

Forensic Value

In directly comparing DNA profiles, there normally is not much point in adding Y-STRs to the autosomal profile. The rarity of the multilocus autosomal STR genotypes already makes them very discriminating.[216] Nevertheless, the patrilineal inheritance of Y chromosomes can be useful for determining the number of male contributors to a mixed semen sample in sexual assault cases; for discerning which male haplotype is present when a genetic male's autosomal STR alleles cannot be detected in the midst of a much larger quantity of female DNA;[217] for familial searching in criminal DNA databases;[218] for inferring biogeographical ancestry as an investigative lead;[219] and for establishing kinship in immigration cases or other situations.[220] A famous example of the last application is the Y-chromosome testing of members of the acknowledged families of President Thomas Jefferson and Sally Hemings that revealed Sally's last child had a fairly rare Y haplotype that President Jefferson—and everyone else in the Jefferson male clan—possessed.[221]

216. Moreover, some geneticists are not confident that Y-STRs are independent of autosomal STRs, making it unclear how to arrive at an estimate for a profile frequency or probability. *See* Weir, *supra* note 182, at 258 ("[A]lthough the dependencies [among autosomal STRs and Y-chromosome and mitochondrial systems] are low, they do exist. Combining match probability estimates over systems is not generally recommended.").

217. *E.g.*, State v. Jones, 345 P.3d 1195 (Utah 2015) (DNA on fingernail clippings from female homicide victim who struggled with her attacker); State v. Maestas, 299 P.3d 892 (Utah 2012) (same).

218. *See infra* section titled "Kinship trawling."

219. *See infra* section titled "Biogeographic ancestry testing."

220. For example, to help identify human remains as those of a male reported as missing, DNA from the remains might be compared to a sample from a father, brother, uncle, or other relative in the same paternal line as the missing person. The presence of the same Y-STRs in both samples tends to show that the remains are indeed those of the missing person.

221. Eugene A. Foster et al., *Jefferson Fathered Slave's Last Child*, 396 Nature 27 (1998), https://doi.org/10.1038/23835. In response to letters emphasizing the inability of markers on the Y chromosome to distinguish between members of the Jefferson clan (including any illegitimate ones, white or black), the authors of the study noted that the genetic information was much more probable if the President rather than one of his sister's children fathered Sally's last child. The paper's title, they acknowledged, was misleading. Eugene A. Foster et al., *The Thomas Jefferson Paternity*

Analytical Methods and Categorical Matches

Typically, the loci used in constructing a Y haplotype for forensic applications are STRs. Up to 30 are in use. For capillary electrophoresis, the same instruments and software used for routine STR testing are employed. Large peaks in an electropherogram that do not seem to be artifacts are interpreted as the alleles of the haplotype.[222] As with autosomal STR profiling, the common practice for comparing a suspect's profile to a trace-evidence one is the two-stage framework of a categorical match or no-match conclusion followed by a statistical evaluation. For instance, the Department of Justice's Uniform Language for Testimony and Reports (ULTR) for Y-STRs states that an examiner "may" testify to an inclusion (meaning "included, or cannot be excluded"), exclusion, or inconclusive (because of an "indistinguishable mixture" or a possible "mutational event").[223] A declared match must be accompanied by some kind of "quantitative statement describing the weight of the evidence."[224]

Case, 397 Nature 32 (1999), https://doi.org/10.1038/16177; *see also* Eliot Marshall, *Which Jefferson Was the Father?*, 283 Science 153 (1999), https://doi.org/10.1126/science.283.5399.153a.

222. Analysts need to be aware of certain quirks with Y-STR profiling to avoid some possible misintepretations of electropherograms. The International Society of Forensic Genetics explains that

> [S]ome mutational effects rarely seen in autosomal STRs are more pronounced in Y-STRs. Especially large-scale deletions, insertions and conversions are responsible for higher numbers of Null . . . and multiple alleles at certain loci. Some markers included in commercial kits show always more than one allele because the sequence has identical copies on the Y chromosome (e.g. DYS385, DYF387S1). One or several Y-STR loci per haplotype can be involved in such mutation events. This is of forensic relevance, because a pattern can be erroneously interpreted as allelic drop-out, DNA contamination[,] and mixture, which may affect the evidential value of a DNA profile.

Lutz Roewer et al., *DNA Comm'n of the Int'l Soc'y of Forensic Genetics (ISFG): Recommendations on the Interpretation of Y-STR Results in Forensic Analysis*, 48 Forensic Sci. Int'l: Genetics 102308, at 2 (2020), https://doi.org/10.1016/j.fsigen.2020.102308 (citations omitted). SWGDAM grandly declares that "[t]he laboratory should establish a method based on validation to document the designation of a peak as an artifact or an allele." Scientific Working Group on DNA Analysis Methods, Interpretation Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories § 4.1 (Mar. 2, 2022), https://perma.cc/ZZR4-FPND. It discusses the difficulties further in related and supplemental documents.

223. U.S. Dep't of Just., Uniform Language for Testimony and Reports for Forensic Y-STR DNA Examinations 2–3 (Dec. 12, 2022) (https://perma.cc/LE6X-FS87) [hereinafter Y ULTR].

224. *Id.* at 3 ("An examiner shall provide a quantitative statement describing the weight of the evidence for all comparisons in which a known male is included as a possible contributor to the Y-STR typing results obtained from a probative evidentiary sample. This statement shall be provided regardless of the number of alleles detected or the magnitude of the resulting quantitative value."). The ISFG is more lenient. Roewer et al., *supra* note 222, at 2 ("Basically, examiners are expected to prepare reports with the minimum requirement of a *qualitative statement* on the Y-STR test result."). *But see* Y ULTR, *supra* note 223, at 5 ("When a suspect is identified that matches the Y-STR profile, we recommend the formulation of hypotheses according to the likelihood approach described by Evett and Weir."); *see also* Ian W. Evett & Bruce S. Weir, Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists (1998).

In court, laboratory performance issues and the manner in which inclusionary findings are presented are more likely to be raised than are any fundamental objections to the science and technology for determining whether two Y-STR profiles are the same. Reasoning that Y-STRs are just another set of markers involving the same principles and procedures as the previously accepted autosomal STRs, courts have upheld the admission of matching Y-STRs.[225] Likewise, actions for postconviction Y-STR testing tend to be resolved on the assumption that this form of DNA testing is admissible.[226]

Population Genetics and Statistics

An exclusion reflects the analyst's belief that differences between the alleles in the profiles are so extensive that they would never be seen in samples within the same paternal lineage. The inference that follows is clear: A correct exclusion means the suspect is not the source. However, the implication of an inclusion or match is not obvious without further information. If the inclusion is correct, the haplotypes are the same. But what follows from that? Does the shared haplotype occur once in a blue moon in the suspect population, or does it pop up all the time? Thus, an inclusion requires a second stage of statistical analysis to know the degree to which the finding, even if correct, tends to prove that the defendant is the source of the trace-evidence sample.[227] The additional information

225. *E.g.,* Shabazz v. State, 592 S.E.2d 876, 879 (Ga. Ct. App. 2004); Commonwealth v. Jacoby, 170 A.3d 1065, 1091–95 (Pa. 2017) (nothing novel about Y-STRs); Curtis v. State, 205 S.W.3d 656, 660–61 (Tex. Ct. App. 2006); Commonwealth v. Clark, 34 N.E.3d 1, 12 (Mass. 2015) (stating "that the results of DNA testing using the Y–STR method are admissible in Massachusetts courts" on the basis of a case with autosomal-STR testing).

226. *E.g.*, United States v. MacDonald, 37 F. Supp. 3d 782 (E.D.N.C. 2014) (motion for Y-STR typing under the Innocence Protection Act of 2004, 18 U.S.C. § 3600, denied as untimely); *Clark*, 34 N.E.3d 1 (denial of a motion for Y-STR testing under a state statute was error).

227. An "inconclusive" means that the DNA test should not be used to decide whether the suspect is the source. It reflects the belief that even if the samples are in the same lineage, the observed differences are not extremely improbable considering the possibility of a mutation or other event. Neither are they very improbable if the samples come from different paternal lines. No inference follows. Yet, DNA analysts sometimes speak of the proportion of a population that would be excluded by haplotype testing as if it were simply 1 minus the proportion of fully matching haplotypes in a reference database for that population. *E.g.*, State v. Polizzi, 924 So.2d 303, 308–09 (La. Ct. App. 2006) (The DNA expert "concluded that the two profiles were consistent with each other, meaning that the Defendant or any of his paternal relatives could not be excluded as having been a donor to the sample from the victim. . . . 99.7 percent of the Caucasian population, 99.8 percent of the African American population, and 99.3 percent of the Hispanic population could be excluded as donors of the DNA in the sample."). But if haplotypes with slightly different STRs—they match except for one mutation—are considered inconclusive, men with those haplotypes cannot be excluded. Hence, the one-off haplotypes should be added in computing the fraction of the population that "could not be excluded." Whether such slightly different haplotypes

needed to evaluate a match comes from population genetics and statistics. To provide this information, researchers have gathered DNA samples from populations both within the United States and across the world and examined them for Y-STRs and Y-SNPs.[228] The goal is to draw on these "reference databases" to estimate the unrelated-man-haplotype frequency or match probability.

Haplotype frequency or probability estimates

Because the entire haplotype (for example, a set of perhaps a dozen to 30 STRs strung out across the Y chromosome) is inherited as a package, it is neither necessary nor appropriate to combine any allele frequencies as is done with autosomal STRs. The haplotype itself is like a single allele, and one can simply count how often it occurs in a sample of unrelated men.[229] As discussed in the *Reference Guide on Statistics and Research Methods*, in this manual, dividing the count (x) by the sample size (n) gives the sample proportion ($p = x/n$). For instance, in *State v. Jones*,[230] a Y-STR haplotype was present in DNA from fingernail clippings from a woman who had been murdered in the Salt Lake City area and who evidently had struggled with her attacker. A laboratory director testified to zero occurrences of the haplotype in a database of size $n = 8{,}028$.[231] The sample proportion in *Jones* was therefore $p = 0/8{,}028 = 0$.

Two variations on the sample proportion x/n as an estimator of a population frequency have some popularity in the forensic-genetics community. Following the testing in the case, the number of times that the haplotype has been seen is no longer x out of n. Now it is $x + 1$ out of the $n + 1$ men with known haplotypes. In *Jones*, this number is 1/8,029, or about 0.012%. Internationally, this "augmented" statistic is typically used as an estimator of the population frequency.[232] There also are arguments for using $(x + 2)/(n + 2)$ or $(x + 2)/(n + 3)$, which would be about 0.025% in *Jones*.[233]

were treated as "consistent with" one another in the percentages recited in *Polizzi* is ambiguous at best.

228. The freely available Y Chromosome Haplotype Reference Database contains data from "more than 1,300 local population samples with more than 270,000 haplotypes in 136 countries." Lutz Roewer, *Using the YHRD Database for Casework Analysis*, ISHI Rep. Apr. 2019, https://perma.cc/25AP-F6VJ.

229. In this context, "unrelated people" means individuals with a different paternal lineage.

230. 345 P.3d 1195 (Utah 2015).

231. *Id.* at 1208 n.42.

232. Roewer et al., *supra* note 222, at 3.

233. *See* Michael Coble et al., *Non-autosomal Forensic Markers, in* Forensic DNA Evidence Interpretation 315, 331 (John Buckleton et al. eds., 2d ed. 2016). The SWGDAM Y chromosome guidelines only state that "[t]he laboratory should determine which of the following methods will be used." SWGDAM, *supra* note 222, § 9.2.2.

These sample proportions do not account for sampling error—the inevitable differences between random samples and the population from which they are drawn. Finding no matches in a sample of 8,028 men hardly means that there would be no other paternal line with a matching haplotype among the hundreds of thousands of men in the Salt Lake City area. If there are any, a different sample might have produced some of them. One solution to this "zero numerator" problem[234] (and to the problem of quantifying sampling error generally) is to supply a confidence interval for plausible values of the population frequency. Testimony that incorporates this concept is common. The laboratory director in *Jones* testified "that 'approximately 99.6 percent of . . . the male population can be excluded' as a contributor of the DNA sample but that Mr. Jones could not be excluded" and that "read another way, the frequency of Mr. Jones's DNA profile 'is equivalent to one in 2681 individuals.' He explained this means that 'every time you test . . . a male [in a different paternal line], the probability of that person having that particular DNA profile is approximately one in 2681.'"[235]

How did the witness get from about 0/8,000 to 1/2,681? The latter figure is an answer to the following question: How large would the population proportion have to be before we would expect to encounter random samples of size 8,028 that are devoid of the haplotype (as in our one database of this size) at least 5% of the time? If the proportion were much larger than 1/2,681, then random samples with no matching haplotype would be rarer. As such, an estimate of 1/2,681 for the population haplotype frequency is not plainly incompatible with the finding that the haplotype is not in the database. It is the upper end of a 95% "confidence interval."[236]

Case law generally accepts the confidence-interval procedure for estimating the frequency of a Y haplotype in reference-population databases.[237] It also rejects arguments that quantitative analyses are unfairly prejudicial; however,

234. Kaye & Stern, *supra* note 12, section titled "Other Situations."

235. State v. Jones, 345 P.3d 1195, 1208 (Utah 2015) (omissions in original).

236. When drawing from a population whose proportion is 1/2,681, the probability of a simple random sample of size 8,028 with no matching haplotypes is 5%. For a population proportion of 1/1,744, the probability of no haplotypes is 1%, so it can be called the upper limit of a 99% confidence interval. A 99% interval may sound better than a 95% interval, but the larger proportion of 1/1,744 is less compatible with the database. It generates random samples that lack the haplotype less often.

The calculations here use the Poisson approximation. Other ways to produce confidence intervals are sensible. Coble et al., *supra* note 233, at 338–39; Kaye & Stern, *supra* note 12, section titled "Other Situations." The SWGDAM guidelines cite one procedure. SWGDAM, *supra* note 222, § 9.2.3.

237. *E.g.*, Commonwealth v. Jacoby, 170 A.3d 1065 (Pa. 2017); State v. Tucker, 920 N.W.2d 680 (Neb. 2018); State v. Jones, 345 P.3d 1195 (Utah 2015); State v. Bander, 208 P.3d 1242, 1255 (Wash. Ct. App. 2009) (noting that defendant made no showing of a scientific controversy); *cf.* Commonwealth v. Lally, 46 N.E.3d 41, 52–53 (Mass. 2016) (confidence interval not required if point estimate is explained clearly).

some opinions admonish trial judges and counsel to avoid exaggerating the power of Y haplotypes for source attribution.[238] The scientific consensus seems to be that the "counting method" and its confidence intervals are overly conservative for rare haplotypes.[239] As such, further refinements have been developed,[240] and population-genetics models that take into account the way haplotypes evolve in reproducing populations may be the most technically defensible basis for estimating frequencies and likelihood ratios.[241]

Likelihood ratios (LRs)

Likelihood ratios (defined in the section titled "Mixtures" later in this reference guide) are recommended for Y haplotypes by the International Society of Forensic Genetics (ISFG),[242] are recognized by SWGDAM,[243] and are advocated for reporting DNA and other forensic-science findings by many scholars and other groups.[244] LRs have the potential advantage of clarifying the weight of the evidence with respect to a range of alternative hypotheses. For example, the contribution of the analysis of the Y haplotypes to the historical debate over the paternity of Sally Hemings's children (see section titled "Forensic Value" above)

238. *Jones*, 345 P.3d at 1249.

239. Roewer et al., *supra* note 222, at 2–3; Coble et al., *supra* note 267, at 330–35.

240. One such adjustment is the "kappa method." Butler, *supra* note 113, at 423–24; Coble et al., *supra* note 267, at 332; Roewer et al., *supra* note 222, at 3.

241. The discrete Laplace method has been said to be the most robust. Coble et al., *supra* note 267, at 333; *see also* Roewer et al., *supra* note 222, at 3 ("established in practice" internationally). Another consideration is the adjustment for population structure when the reference database comes from an amalgam of preferentially reproducing subgroups in a larger population (see *supra* section titled "Could an Unrelated Person Be the Source?"). For advice, see Coble et al., *supra* note 233, at 329, 330, 334–35; Roewer et al., *supra* note 222, at 5.

242. Roewer et al., *supra* note 222, at 5. The ISFG gives a generic example of possible wording to explain the number:

> The Y-STR profile detected in the crime stain is LR times more probable to observe under hypothesis H_1 than under hypothesis H_2. This notwithstanding, paternal relatives have a high probability to have the same Y-STR profile and will in that case have the same likelihood ratio (LR).

Id.

243. SWGDAM, *supra* note 222, § 9.2.5 ("The laboratory should determine if likelihood ratios will be used to provide quantitative assessments of the value of the matches using relevant populations.").

244. *See* section titled "Frequencies, Probabilities, and Prejudice" above. The Department of Justice ULTR demands that "[a]n examiner shall provide a quantitative statement describing the weight of the evidence for all comparisons in which a known male is included as a possible contributor to the Y-STR typing results obtained from a probative evidentiary sample." Y ULTR, *supra* note 223, at 3. Presumably, the LR can be the "quantitative statement [of] weight" for inclusions, but the ULTR does not address its use for exclusions (where the LR is very large for the hypothesis that a man other than the defendant is the source as opposed to the hypotheses that the defendant is the source) and inconclusives (where the LR is 1).

might have been clearer from the outset had the researchers made statements for each hypothesis about paternity—for example:

- The Jefferson haplotype is at least 100 times more probable if the president was the father of Eston Hemings Jefferson than if one of the president's sister's boys was the father;
- The Jefferson haplotype has the same probability if the president was the father as it does if the president's brother was the father.

Similarly, in *State v. Jones*,[245] the likelihood approach to reporting results would transform the testimony that "this means that 99.92 percent of the male population could be excluded as a possible donor" into:

- The haplotype is about 2,680 times more probable if Mr. Jones is its source than if any particular man not in Mr. Jones's paternal family tree is the source; and
- The haplotype has the same probability if Mr. Jones is its source as it does if any other man in his paternal family tree (a father, child, uncle, certain cousins, nephews, etc.) is the source.

In addition to its clarity with regard to hypotheses in the case, the broader likelihood ratio framework does not require the match/no-match categories and rules for "inconclusives." It can better describe the evidentiary value of haplotypes that are one mutation away from a clear match[246] and can be applied to ambiguous data from low-template DNA.[247] However, questions can remain as to the computation of likelihood ratios in these more challenging cases.

Mitochondrial DNA

Mitochondria and Their Genomes

Mitochondria are small structures, with their own membranes, found inside the cell but outside its nucleus. Within these organelles, molecules are broken down to supply energy. Mitochondria have a small genome—a circle of 16,569 nucleotide base pairs that includes only 37 genes. They started out as bacteria that were engulfed by cells billions of years ago. Many of their original genes migrated to

245. *See supra* section titled "Genetic Principles."

246. Coble et al., *supra* note 233, at 340; *cf. supra* note 227 (on the impact of "inconclusives" on exclusion probabilities).

247. Coble et al., supra note 233, at 342–44.

the nuclear chromosomes,[248] where they produce proteins and RNAs for the mitochondria.[249] However, the sequences currently found in the mitochondria themselves bear little relation to the comparatively monstrous chromosomal genome in the cell nucleus. In particular, they are not physically or statistically associated with the forensic STRs.

Forensic Value

Mitochondrial DNA (mtDNA) has four features that make it useful for forensic DNA testing. First, the typical cell, which has but one nucleus, contains hundreds or thousands of nearly identical mitochondria. Hence, for every copy of chromosomal DNA, there are hundreds or thousands of copies of mitochondrial DNA. This makes it possible to detect mtDNA in samples, such as bone and hair shafts, that contain too little nuclear DNA for conventional typing.

Second, two hypervariable regions that tend to be different in different individuals lie within the control region or D-loop (displacement loop) of the mitochondrial genome.[250] These regions extend for a bit more than 300 base pairs each—short enough to be typable even in highly degraded samples such as very old human remains.

Third, mtDNA comes solely from the egg cell.[251] For this reason, mtDNA is inherited maternally, with no paternal contribution:[252] Full siblings, maternal half-siblings, and others related through maternal lineage normally possess the same mtDNA sequence. This feature makes mtDNA particularly useful for

248. James B. Stewart & Patrick F. Chinnery, *Extreme Heterogeneity of Human Mitochondrial DNA from Organelles to Populations*, 22 Nature Rev. Genetics 106, 106–07 (2021), https://doi.org/10.1038/s41576-020-00284-x.

249. Whole-genome sequencing of parents and their children establishes that other segments of mitochondrial DNA (mtDNA) continue to find their way into the nuclear genome, creating new nuclear mtDNA segments (NUMTs) in some cells. *E.g.*, Wei Wei et al., *Nuclear-embedded Mitochondrial DNA Sequences in 66,083 Human Genomes*, 611 Nature 105 (2022), https://doi.org/10.1038/s41586-022-05288-7.

250. A third, somewhat less polymorphic region in the D-loop, can be used for additional discrimination. The remainder of the control region, although noncoding, consists of DNA sequences that are involved in the transcription of the mitochondrial genes. These control sequences are essentially the same in everyone (monomorphic).

251. The relatively few mitochondria in the spermatozoan that fertilizes the egg cell soon degrade and are not replicated in the multiplying cells of the pre-embryo.

252. The possibility of paternal contributions to mtDNA in humans is discussed in Alistair T. Pagnamenta et al., *Biparental Inheritance of Mitochondrial DNA Revisited*, 22 Nature Rev. Genetics 477 (2021), https://doi.org/10.1038/s41576-021-00380-6 ("Evidence for a biparental mode of . . . inheritance has been sparse and remains controversial. Recent studies using a range of complementary techniques do not support paternal transmission of mtDNA and highlight the co-amplification of rare, concatenated nuclear mtDNA segments as a technical artefact that may explain previous observations.").

associating persons related through their maternal lineage. It has been exploited to identify the remains of the last Russian tsar and other members of the royal family, of soldiers missing in action, and of victims of mass disasters.[253]

Finally, point mutations accumulate, particularly in the noncoding D-loop, without altering how the mitochondrion functions. Hence, a single individual can develop distinct internal populations of mitochondria. Single nucleotide and other variants can occur within a cell, between cells in a given tissue, and between cells from different tissues and organs.[254] Such "heteroplasmy" has been observed in healthy humans, with an average of one heteroplasmic site per individual.[255] As discussed below, heteroplasmy complicates the interpretation of differences in mtDNA sequences.[256] Yet, it is mutations that make mtDNA polymorphic and hence useful in identifying individuals. Over time, mutations in egg cells can propagate to later generations, producing more heterogeneity in the inherited mitochondrial genomes in the human population.[257] This polymorphism allows scientists to compare mtDNA from crime scenes to mtDNA from given individuals to ascertain whether the tested individuals are within the maternal line (or another coincidentally matching maternal line) of people who could have been the source of the trace evidence.

Analytical Methods and Categorical Matches

The small mitochondrial genome can be analyzed with sequencing methods or with microarrays that give the order of some or all the base pairs.[258] The traditional Sanger-sequencing technology usually was done on just the hypervariable parts of the coding region. Cheaper and faster massively parallel sequencing is replacing that procedure, increasing the ability of laboratories to detect heteroplasmy and better distinguish among different maternal lineages in the population.[259] The laboratory first generates descriptions of the sequences of two

253. *See, e.g.,* Silent Witness: Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters (Henry Ehrlich et al. eds., 2020).

254. "Length heteroplasmy" (most often in the form of simple repeats such as CC . . .) is common but harder to characterize precisely in sequencing studies.

255. Alison Barrett et al., *Pronounced Somatic Bottleneck in Mitochondrial DNA of Human Hair*, 375 Phil. Trans. Roy. Soc'y B 20190175, at 2 (2019), https://doi.org/10.1098/rstb.2019.0175.

256. *See infra* section titled "Population Genetics and Statistics."

257. Estimates of the average mutation rate range up to around 1 per 70 generations. Mikkel M. Andersen & David J. Balding, *How Many Individuals Share a Mitochondrial Genome?,* 14 PLoS Genetics e1007774 (2018), https://doi.org/10.1371/journal.pgen.1007774.

258. *See supra* sections titled "Sequence-Specific Probes and Microarrays" and "Sequencing and SNPs."

259. For a survey of the major methods, see Bethany Forsythe et al., *Methods for the Analysis of Mitochondrial DNA*, 3 WIREs Forensic Sci. e1388 (2021), https://doi.org/ 10.1002/wfs2.1388.

samples—say, DNA extracted from a hair shaft found at a crime scene and hairs plucked from a suspect.[260] Let's assume that there is no issue as to whether the technology has been properly applied with suitable controls,[261] and the factfinder can be confident that the sequences as reported truly represent the order of the base pairs in the mtDNA of the hairs. What should the factfinder make of the two sequences?

Most analysts use match/no-match categories to describe the result of a comparison.[262] The Department of Justice currently favors the words "[c]annot be excluded (i.e., inclusion, or included) [and] [e]xclusion (i.e., excluded)."[263] As will be explained in the section below titled "Heteroplasmy," an intermediate category of "inconclusive" covers less-than-perfectly-matching sequences.[264] In the simplest cases, the two sequences show a good number of differences (a clear exclusion), or they are identical (an inclusion). Suppose there is an exclusion. After explaining why a clear "exclusion" means that such a large difference almost certainly cannot occur for two hairs from people within the same lineage, an analyst's testimony on direct examination could end. But if the analyst were to stop at the same point when declaring an inclusion, the factfinder would be left with no scientific information on how common or rare such matching sequences are for different lineages in the relevant population. Such inclusion-only testimony invites the objection that the congruent sequences are of unknown probative value and might be given more weight than they deserve.[265] To more

260. Mitotyping often can exclude individuals as the source of stray hairs even when the hairs are microscopically indistinguishable. Nonetheless, gross or microscopic hair comparison can be useful, as a screening test, to exclude a suspect without the need for DNA analysis. *See* ASTM E3316–22, Standard Guide for Forensic Examination of Hair by Microscopy (2022).

261. *See* Walther Parson et al., *DNA Comm'n of the Int'l Soc'y for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing*, 13 Forensic Sci. Int'l: Genetics 134, 135–36 (2014), https://doi.org/10.1016/j.fsigen.2014.07.010; Scientific Working Group on DNA Analysis Methods, SWGDAM Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories § 1 (Apr. 23, 2019) [hereinafter SWGDAM MtDNA Guidelines].

262. *E.g.*, Lewis v. State, 889 So.2d 623, 673 (Ala. Ct. Crim. App. 2003) (FBI expert testified that "[t]he fourth step is to compare the order of the bases, i.e., the sequence, to other samples to determine if there is a 'match.'"); *see also* SWGDAM MtDNA Guidelines, *supra* note 261, § 3.1.

263. U.S. Dep't of Just., Uniform Language for Testimony and Reports for Forensic Mitochondrial DNA Examinations 2 (Dec. 12, 2022), https://perma.cc/J6V4-Z4EG [hereinafter MtDNA ULTR]. This terminology document does not single out "match" or words based on that stem as impermissible. It defines these labels as "an examiner's conclusion" with no specification for how the examiner should reach these conclusions.

264. In principle, a likelihood ratio is better suited to cases in which there are slight sequence differences (and could be used in all situations). *See* Coble et al., *supra* note 233, at 319, 324–41; *cf. supra* section titled "Likelihood ratios (LRs)" (on likelihood ratios for Y-STRs).

265. *But see* Commonwealth v. Chmiel, 30 A.3d 1111, 1158 (Pa. 2011) (suggesting that "statistical analysis is . . . unnecessary").

adequately inform the court or jury of the implications of a match, forensic scientists invoke principles of population genetics and statistics.[266]

Population Genetics and Statistics for Mitochondrial DNA

Population databases

As with Y-STRs, to indicate the significance of the match, analysts usually estimate the frequency of the sequence in some population. It is not necessary to combine any allele frequencies because the entire mtDNA sequence, whatever its internal structure may be, is inherited as a single unit (a haplotype). In other words, the sequence itself is like a single allele, and one can simply count how often it occurs in a sample of unrelated people.[267]

Laboratories therefore refer to databases of mtDNA sequences to see how often the type in question has been seen before and to compute the probability of a matching haplotype in a random draw from the population given that it has been observed in the case under investigation.[268] The issues are essentially the same as those for Y haplotypes.[269] Key questions are which reference population or populations to use;[270] what estimator to use for the sequence frequency or match probability;[271] how to account for sampling error in this estimate;[272]

266. The Department of Justice expects analysts in its laboratories to "provide a quantitative statement describing the weight of the evidence for all inclusions regardless of the magnitude of the resulting quantitative value." MtDNA ULTR, *supra* note 263, at 3.

267. In this context, "unrelated people" means individuals with a different maternal lineage.

268. The publicly available European DNA Profiling Group (EDNAP) mtDNA Population Database (EMPOP), which houses quality-controlled sequence data from populations across the world, was established in 2006. Walther Parson & Arne Dür, *EMPOP—A Forensic mtDNA Database*, 1 Forensic Sci. Int'l: Genetics 88 (2007), https://doi.org/10.1016/j.fsigen.2007.01.018.

269. *See supra* section titled "Population Genetics and Statistics."

270. *See* Parson et al., *supra* note 261, at 140 ("Local, regional and continental databases may all be searched and reported to guide evaluation of the evidence. Additionally, frequencies from geographic databases of relevant differentiated populations may be reported (as in the United States where reports often list separately search results from major population groups)."); *supra* section titled "Haplotype frequency or probability estimates."

271. *See* Parson et al., *supra* note 261, at 140; *supra* section titled "Haplotype frequency or probability estimates" (noting most of the alternatives). In *State v. Pappas*, 776 A.2d 1091 (Conn. 2001), the defendant maintained that the usual sample proportion is inadmissible because an estimate that tosses in the haplotype in the case itself is more appropriate. The Connecticut Supreme Court responded that "using zero as the numerator rather than one . . . goes to the weight of the evidence and not to its admissibility." *Id.* at 1109.

272. *See* Parson et al., *supra* note 261, at 140 ("approaches that empirically take into account the haplotype distribution of the database in question (e.g. the Kappa model . . .) may best represent the strength of the mtDNA evidence"); SWGDAM MtDNA Guidelines, *supra* note 261, § 4.2.4 (binomial distribution); *supra* section titled "Haplotype frequency or probability estimates"; *infra* note 275.

whether and how to adjust for population structure;[273] whether to present likelihood ratios, including one that combines the information in mitochondrial haplotypes with that from Y haplotypes (or both Y and autosomal haplotypes);[274] and how to present or explain the statistical assessment to a factfinder.[275]

Heteroplasmy

The simple inclusion–exclusion approach must be modified to account for the fact that the same individual can have detectably different (albeit very similar) mitotypes in different tissues or even in different cells in the same tissue. To understand the implications of heteroplasmy, we need to understand how it comes into existence. Heteroplasmy can occur because of mtDNA mutations during the division of adult cells, such as those at the roots of hair shafts. These new mitotypes are confined to the individual. They will not be passed on to future generations. Heteroplasmy also can result from a mutation contained in the egg cell that grew into an individual. Such mutations can make their way into succeeding generations, establishing new mitotypes in the population. But this is an uncertain process. Egg cells contain many mitochondria, and the

273. *See supra* section titled "Haplotype frequency or probability estimates." *Compare* Coble et al., *supra* note 233, at 334–35 (describing a procedure) *with* SWGDAM MtDNA Guidelines, *supra* note 261, § 4.3 ("However, determination of an appropriate theta (θ) value is complicated by the variety of primer sets, covering different portions of HV1 and/or HV2, which may be applied to forensic casework. SWGDAM has not yet reached consensus on the appropriate statistical approach to estimating θ for mtDNA comparisons.").

274. *See* Parson et al., *supra* note 261, at 140; SWGDAM MtDNA Guidelines, *supra* note 261, §§ 4.4 & 4.5.

275. Apparently, SWGDAM would dispense with estimates of sampling error. SWGDAM MtDNA Guidelines, *supra* note 261, § 4.2.3 ("Reporting an mtDNA haplotype sample frequency without a confidence interval is acceptable as a factual statement regarding observations in the database."). But the conclusion that judges or jurors can make suitable use of a sample proportion without an assessment of statistical uncertainty is not a strictly scientific judgment. A court might need to consider whether it is sufficient to leave it to the jury to decide how to weigh the fact of the match and the absence of the same sequence in a convenience sample that might—or might not—be representative of the local population.

As with Y haplotyping (*supra* note 227), the use of exclusion probabilities that do not account for inconclusives could be misleading. In *Vaughn v. State*, 646 S.E.2d 212, 215 (Ga. 2007), for instance, the statement that a suspect "cannot be excluded" when "there is a single base pair difference" was transformed into "a match." These inconclusive sequences contribute to the number of people who would not be excluded. Therefore, in *Pappas*, it was misleading to conclude "that approximately 99.75% of the Caucasian population could be excluded as the source of the mtDNA in the sample." 776A.2d 1091, 1104 (Conn. 2001) (footnote omitted). This percentage neglects the individuals whose mtDNA sequences are off by one base pair. Along with the 0.25% who are included because their mtDNA matches completely, these one-off people would not be excluded. Of course, the difference may be fairly small. In *Pappas*, a defense expert reported that the actual nonexclusion rate was still "99.3 percent of the Caucasian population." *Id.* at 1105 (footnote omitted).

mature egg cell will not contain just the mutation—it will house a mixed population of the old-style mitochondria and a number of the mutated ones (with DNA that usually differs from the original at a single base pair). Figuratively speaking, the original mtDNA sequence and the mutated version fight it out for several generations until one of them becomes "fixed" in the population. In the interim, the progeny of the mutated egg cell will harbor both strains of mitochondria.

When mtDNA from a single-source crime-scene sample is compared to a suspect's sample, there are three possibilities: (1) neither sample is detectably heteroplasmic; (2) one sample displays heteroplasmy, but the other does not; (3) both samples display heteroplasmy. In each scenario, the comparison can produce an exclusion or an inclusion:

1. *Neither sample heteroplasmic.* In the first situation, if the sequence in the crime-scene sample is markedly different from the sequence in the suspect's sample, then the suspect is excluded. But heteroplasmy could be the reason for a difference of only a single base or so. For example, the sequence in a hair shaft coming from the suspect could be a slight mutation of the dominant sequence in the suspect. Therefore, the FBI treats a difference at a single base pair as inconclusive.[276] When the one mtDNA sequence characteristic of each sample is identical, the issue becomes how to use the reference database of mtDNA sequences, as discussed above.
2. *Suspect's sample heteroplasmic, crime-scene sample not.* One version of the second scenario arises when heteroplasmy is seen in the suspect's tissues but not in the crime-scene sample. If the crime-scene sequence is not close to either of the suspect's sequences, then the suspect is excluded. If it is identical to one of the suspect's sequences, then the suspect is included, and a suitable reference database should indicate how infrequent such an inclusion would be. If crime-scene DNA is one base pair removed from either of the suspect's sequences, then the result is inconclusive.
3. *Both samples heteroplasmic.* In this third scenario, multiple sequences are seen in each sample. To keep track of things, we can call the sequences in the crime-scene sample C1 and C2, and those in the suspect's sample S1 and S2. If either C1 or C2 is very different from both S1 and S2,

276. SWGDAM MtDNA Guidelines, *supra* note 261, § 3.1. Where did the one-base-pair rule come from? With traditional Sanger sequencing of parts of the control region of the mitogenome, heteroplasmy at a single point seems to occur, on average, in approximately 6% of the control-region sequences from around the world; heteroplasmy at multiple sites is seen in less than 1% of sequences. Jodi A. Irwin et al., *Investigation of Heteroplasmy in the Human Mitochondrial DNA Control Region: A Synthesis of Observations from More Than 5000 Global Population Samples*, 68 J. Molecular Evolution 516 (2009), https://doi.org/10.1007/s00239-009-9227-4.

the suspect is excluded. If C1 and C2 are the same as S1 and S2, the suspect is included. Because detectable heteroplasmy is not very common, this inclusion is stronger evidence of identity than the simple match in the first scenario. Finally, in some of the situations in which C1 or C2 are merely very close to S1 or S2, the comparison will be inconclusive.

Emerging Questions for Courts

A number of courts have rejected objections that Sanger sequencing of hypervariable regions of the mitochondrial DNA genome does not comport with *Frye*[277] or *Daubert*.[278] But forensic-DNA laboratories are adopting some of the massively parallel sequencing (MPS) methods, described in the section above titled "Sequencing methods," that can work with smaller quantities of mtDNA and that can efficiently sequence the entire mitogenome.[279] The widespread use of MPS methods in clinical medicine and genetic research indicates scientific acceptance and validity at a general level (see "Sequencing methods"), but courts still may encounter questions about the rate of errors in the "reads," what steps are taken to correct them, precautions against misinterpreting nuclear mtDNA segments (NUMTs) as present in mitochondria, and more. Much depends on the equipment and its operation.[280] Published validity studies from biotechnology companies developing the sequencers and kits, from the individual forensic

277. *E.g.,* Magaletti v. State, 847 So. 2d 523, 528 (Fla. Dist. Ct. App. 2003) ("[T]he mtDNA analysis conducted [on hair] determined an exclusionary rate of 99.93 percent. In other words, the results indicate that 99.93 percent of people randomly selected would not match the unknown hair sample found in the victim's bindings."); People v. Sutherland, 860 N.E.2d 178, 271–72 (Ill. 2006); People v. Holtzer, 660 N.W.2d 405, 411 (Mich. Ct. App. 2003); Wagner v. State, 864 A.2d 1037, 1043–49 (Md. Ct. Spec. App. 2005) (mtDNA sequencing admissible despite contamination and heteroplasmy).

278. *E.g.*, United States v. Beverly, 369 F.3d 516, 531 (6th Cir. 2004) ("The scientific basis for the use of such DNA is well established."); United States v. Coleman, 202 F. Supp. 2d 962, 967 (E.D. Mo. 2002) ("'[a]t the most,' seven out of 10,000 people would be expected to have that exact sequence of As, Ts, Cs, and Gs"), *aff'd*, 349 F.3d 1077 (8th Cir. 2003); *Pappas*, 776 A.2d at 1095; State v. Underwood, 518 S.E.2d 231, 240 (N.C. Ct. App. 1999); State v. Council, 515 S.E.2d 508, 518 (S.C. 1999); State v. Griffin, 384 P.3d 186, 202 (Utah 2016).

279. The methods being implemented are not the most advanced ones that have been proposed for mitochondrial research. It is technically possible to zoom in on the mitogenome (as a single molecule, without PCR amplification) and sequence it in a single "read." Ieva Keraite et al., *A Method for Multiplexed Full-Length Single-Molecule Sequencing of the Human Mitochondrial Genome*, 13 Nature Communications 5902 (2022), https://doi.org/10.1038/s41467-022-33530-3 (using CRISPR-Cas9 and claiming to overcome the relatively large rate of read errors in single-molecule sequencing systems).

280. *See, e.g.,* Yun Heo et al., *Comprehensive Evaluation of Error-Correction Methodologies for Genome Sequencing Data, in* Bioinformatics (Nakaya Helder ed., 2021); Nicholas Stoler & Anton

laboratories deploying them, and from academic researchers can be scrutinized.[281] At least in early admissibility hearings, courts may wish to consider appointing independent expert advisers for this purpose.

Questions of the accuracy of the sequences and their interpretation also might emerge with respect to the "reliably applied" prong of the requirements for scientific evidence under Rule 702. The presence of and adherence to rigorous quality control and assurance systems might be significant.[282] Moreover, all these matters are potential topics that can arise for jurors (and judges when they act as triers of fact). The practice followed in a case at bar can be compared to both minimally acceptable and recommended practices as articulated in the scientific literature and guidelines from expert groups.[283]

Courts have consistently held that neither the phenomenon of heteroplasmy nor the limitations in the statistical analysis preclude the forensic use of sequencing results under either Rule 702 or Rule 403.[284] However, mitogenome MPS picks up more heteroplasmic sites than Sanger sequencing, and more individuals exhibit not just one, but several points of heteroplasmy.[285] As a result, restricting the "inconclusive" category to only a single sequence-difference across the mitogenome will not prevent as many false exclusions. Despite the allure of purportedly definitive "exclusions" and "inclusions," the gray zone of "inconclusive" sequence differences is increasingly hard to define with a simple rule of thumb for the entire mitogenome. Courts may need to inquire into what the literature demonstrates about the accuracy of the laboratory's rule or procedure for assessing the extent to which sequence similarities and differences support conclusions (including judgments of "inconclusive") about the maternal relatedness of the unknown individual whose mtDNA was recovered and the suspect's comparison sample.

Nekrutenko, *Sequencing Error Profiles of Illumina Sequencing Instruments*, 3 Nucleic Acids Rsch. Genomics & Bioinformatics lqab019 (2021), https://doi.org/10.1093/nargab/lqab019.

281. *E.g.*, Michael D. Brandhagen et al., *Validation of NGS for Mitochondrial DNA Casework at the FBI Laboratory*, 44 Forensic Sci. Int'l: Genetics 102151 (2020), https://doi.org/10.1016/j.fsigen.2019.102151 (sequencing the control region); Jennifer Churchill Cihlar et al., *Developmental Validation of a MPS Workflow with a PCR-Based Short Amplicon Whole Mitochondrial Genome Panel*, 11 Genes 1345 (2020), https://doi.org/10.3390/genes11111345 (whole-mitogenome sequencing).

282. *See supra* section titled "Sample Collection and Laboratory Performance."

283. *See supra* note 261.

284. *E.g.*, *Beverly*, 369 F.3d at 531 ("[T]he mathematical basis for the evidentiary power of the mtDNA evidence was carefully explained, and was not more prejudicial than probative."); *Pappas*, 776 A.2d 1091; *Griffin*, 384 P.3d at 202–03.

285. Jennifer A. McElhoe et al., *Exploring Statistical Weight Estimates for Mitochondrial DNA Matches Involving Heteroplasmy*, 136 Int'l J. Legal Med. 671, 695–96 (2022), https://doi.org/10.1007/s00414-022-02774-5.

Mixtures

What Are DNA Mixtures?

Samples of biological trace evidence recovered from crime scenes often contain a mixture of fluids or tissues from different individuals. Examples include vaginal swabs collected as sexual assault evidence, bloodstain evidence from scenes where several individuals shed blood, and objects that several people touched. However, not all mixed samples produce mixed STR profiles. Consider a sample in which 99% of the DNA comes from the defendant and 1% comes from a different individual. Even if some of the molecules from the minor contributor come in contact with a primer and the polymerase and an STR is amplified, the resulting signal might be too small to detect—the peak in an electropherogram will blend into the background. Because the vast bulk of the amplified STRs will come from the defendant's DNA, the electropherogram may show only one STR profile. In these situations, the interpretation of the single DNA profile is the same as when 100% of the DNA molecules in the sample are the defendant's.

When the mixtures are more evenly balanced among contributors, however, the STRs from multiple contributors can appear as "extra" peaks. As a rule, because DNA from a single individual can have no more than two alleles at each locus,[286] the presence of three or more peaks at several loci indicates a mixture of DNA in the sample.[287] Figure 6 shows another electropherogram from DNA recovered in *People v. Pizarro*.[288] The fact that there are as many as four alleles at

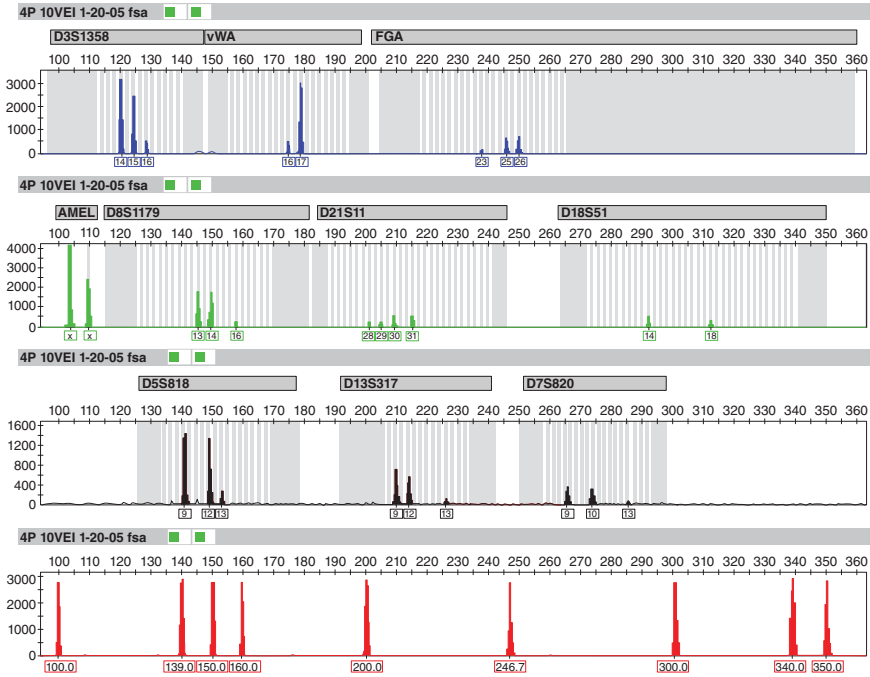
286. This follows from the fact that individuals inherit chromosomes in pairs, one from each parent. *See supra* section titled "Sexual Reproduction and the Genome." An individual who inherits the same allele from each parent (a homozygote) can contribute only that one allele to a sample, and an individual who inherits a different allele from each parent (a heterozygote) will contribute those two alleles. Finding three or more alleles at several loci therefore indicates a mixture.

287. On rare occasions, an individual exhibits a phenotype with three alleles at a locus. This can be the result of a chromosome anomaly (such as a duplicated gene on one chromosome or a mutation). A sample from such an individual is usually easily distinguished from a mixed sample. The three-allele variant is seen at only the affected locus, whereas with mixtures, more than two alleles typically are evident at several loci.

Individuals who are genetic chimeras also can produce DNA samples that could be mistaken for that of two different people. Chimerism refers to more than one cell line resulting from different zygotes (fertilized eggs). Chimerism can arise artificially, as a result of a blood transfusion or transplantation, or spontaneously, during the formation and development of an embryo. Chimerism: A Clinical Guide 3–48 (Nicole L. Draper ed., 2018). The implications for forensic-DNA identification are discussed in, e.g., David H. Kaye, *Chimeric Criminals*, 14 Minn. J.L. Sci. & Tech. 1 (2012); Erin Murphy, *Inside the Cell: The Dark Side of Forensic DNA* 246–49 (2015); Kayla Sheets & Robert Wenk, *Relationship Testing and Forensics*, *in* Chimerism: A Clinical Guide, *supra*, at 51–63.

288. *See supra* Figure 4.

Figure 6. Electropherogram in *People v. Pizarro* that can be interpreted as a mixture of DNA from the victim and the defendant.



Source: Steven Myers and Jeanette Wallin, California Department of Justice, provided the image.

That some loci in Figure 6, such as vWA, exhibit only two peaks does not preclude the existence of a two-person mixture. Both the victim and the rapist (on the state's theory of the case) might share the same alleles at that locus, and the amplified product of the victim's DNA might be masking that from the rapist. Both might be homozygous at that locus (each producing one peak). If it is credible that one or both alleles of the rapist's vWA locus dropped out of the electropherogram (not likely with substantial quantities of DNA), still more scenarios consistent with the state's theory of the case could explain why there are only two peaks at the vWA locus. Enumerating and assessing such possibilities leads to the topic of "mixture deconvolution."

some loci and that many of the peaks match the victim's suggests that the sample is a mixture of the victim's and another person's DNA. Furthermore, the presence of two peaks at the amelogenin locus (labeled AMEL in the electropherogram) shows that male DNA is part of the mixture. Because all the peaks that do not match the victim are part of the defendant's STR profile, the mixture is consistent with the state's theory that the defendant raped the victim.

Have Strategies for Simplifying Mixtures Interpretation Been Pursued?

A pioneer in the development of DNA typing methods and their interpretation had this advice for forensic laboratories: "Don't do mixture interpretations unless you have to."[289] If a laboratory has other samples that do not show evidence of mixing, it may be able to avoid fully deciphering mixed profiles. Even across a single stain, the proportions of a mixture can vary, and it might be possible to extract a DNA sample that does not produce a mixed profile.

In sexual assault cases, it sometimes is possible to separate sperm from the male and female epithelial cells on vaginal, oral, or rectal swabs and then analyze the sperm DNA by itself.[290] When this procedure works well, and only one man's sperm are present, the electropherogram should display one clear profile with little or no interference from the female DNA. And even with sperm from more than one man, the analysis might be simpler because the victim's DNA no longer can mask alleles from the sperm. Also, in sexual assault (and other) cases, Y chromosome testing can reveal the number of men (from different paternal lines) whose DNA is being detected and whether the defendant's Y chromosome is consistent with one of these paternal lines.[291] Because only males have Y chromosomes, the female DNA in a mixture has no effect.

Finally, rather than extracting DNA from a mixture of cells and performing PCR-CE on the resulting mixture of DNA molecules, it may be possible to isolate individual cells and analyze the DNA one cell at a time,[292] either by pushing

289. Peter Gill, as quoted in John M. Butler, *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* 166 (2d ed. 2005).

290. The nucleus of a sperm cell lies behind a protective structure that does not break down as easily as the membrane in an epithelial cell. This makes it possible to disrupt the male and female epithelial cells first and remove their DNA, and then to use a harsher treatment to disrupt the sperm cells. Heather E. McKiernan & Phillip B. Danielson, *Molecular Diagnostic Applications in Forensic Science*, *in* Molecular Diagnostics 371 (George P. Patrinos ed., 3d ed. 2017). Other methods of differential extraction also have been studied or applied to forensic samples. *E.g.*, Lana Ostojic et al., *Micromanipulation of Single Cells and Fingerprints for Forensic Identification*, 51 Forensic Sci. Int'l: Genetics 102430 (2021), https://doi.org/10.1016/j.fsigen.2020.102430.

291. *E.g.*, State v. Polizzi, 924 So. 2d 303, 308–09 (La. Ct. App. 2006) (testing for Y-STRs on "the genital swab with the DNA profile from the Defendant's buccal swab, . . . the Defendant or any of his paternal relatives could not be excluded as having been a donor to the sample from the victim," while "99.7 percent of the Caucasian population, 99.8 percent of the African American population, and 99.3 percent of the Hispanic population could be excluded as donors of the DNA in the sample"); State v. Bander, 208 P.3d 1242, 1250 (Wash. Ct. App. 2009) (Y-STR profiling showed that "99.9 percent of the population could be excluded as possible contributors to the sample extracted from the electrical cords recovered from Gardner's body").

292. *See* Jianye Ge et al., *Precision DNA Mixture Interpretation with Single-Cell Profiling*, 12 Genes 1649 (2021), https://doi.org/10.3390/genes12111649.

PCR-CE to its limits[293] or by employing the newer sequencing methods sketched in the section above titled "Sequencing methods."[294] Without physical separation, however, the multiple genotypes in the sample can produce mixed profiles. Analysts then have to interpret these more complex patterns in terms of the individual genotypes that might be present. Inferring the probable genotypes that are combined in a mixed profile is known as mixture deconvolution. This interpretive process is described in the next subsection.

What Is Mixture Deconvolution?

The goal of mixture deconvolution is twofold: (1) to deduce the individual genotypes that are intertwined in an electropherogram that contains peaks from the DNA of different individuals; and (2) to assess the significance of the fact that a defendant has some or all of such a genotype.[295] This analysis is computational. A simplified example illustrates the nature of the procedure that has been in use for over 20 years.[296] Suppose we create a mixture by adding equal, easily detectable amounts of DNA from two unrelated individuals and type it at just one autosomal STR locus. Let us say that the contributors are each heterozygotes—each individual has two alleles per locus—and that all the alleles are distinct. The electropherogram will have four peaks of roughly equal heights at locations that we can designate A, B, C, and D. An analyst who is given this electropherogram without any information about the contributors could reason as follows:

293. *E.g.*, Kaitlin Huffman et al., *Y-STR Mixture Deconvolution by Single Cell Analysis*, 68 J. Forensic Sci. 275 (2023), https://doi.org/10.1111/1556-4029.15150 (individual Y-STR profiles recovered from direct PCR amplification of single-cell subsamples from two- and three-person mixtures).

294. *See also* section titled "Did the Sample Contain Enough DNA?" and note 112 above; *cf.* Lucie Kulhankova et al., *Single-Cell Transcriptome Sequencing Allows Genetic Separation, Characterization and Identification of Individuals in Multi-Person Biological Mixtures*, 6 Communications Biology 201 (2023), https://doi.org/10.1038/s42003-023-04557-z (initial study using single-cell RNA transcript sequences and other genetic data to successfully deconvolute laboratory-created mixtures of blood from 2- to 9-person mixtures in ratios as low as 1 to 60).

295. Some courts have held mixture interpretations admissible without any statistical assessment. *E.g.*, Rodriguez v. State, 273 P.3d 845, 851–52 (Nev. 2012); People v. Her, 157 Cal. Rptr. 3d 40 (Cal. Ct. App. 2013).

296. Tim M. Clayton et al., *Analysis and Interpretation of Mixed Forensic Stains Using DNA STR Profiling*, 91 Forensic Sci. Int'l 55 (1998), https://doi.org/10.1016/s0379-0738(97)00175; Ian W. Evett et al., *Taking Account of Peak Areas When Interpreting Mixed DNA Profiles*, 43 J. Forensic Sci. 62 (1998); Peter Gill et al., *Interpreting Simple STR Mixtures Using Allele Peak Areas*, 91 Forensic Sci. Int'l 41 (1998), https://doi.org/10.1016/s0379-0738(97)00174-6.

The minimum number of contributors is two, but there could be more. The four peaks are of equal height, so if there are only two contributors to the mixture, their genotypes could be any of the following six combinations:

CONTRIBUTOR 1	CONTRIBUTOR 2
AB	CD
AC	BD
AD	BC
CD	AB
BD	AC
BC	AD

These can be written more compactly using a slash to separate the two contributors' genotypes and listing them as AB/CD, AC/BD, etc. There cannot be exactly three contributors (unless one of them did not contribute enough DNA to have any impact on the peak heights).[297] If there are four contributors, the genotypes could be AB/CD/AB/CD, AC/BD/AC/BD, AD/BC/AD/BC, AA/BB/CC/DD, AC/BC/AB/DD, AC/BC/AD/BD, AC/AD/CD/BB, or BC/BD/CD/AA and more (rearranging the order to assign the two-allele genotypes to the different contributors 1 through 4). As five or more contributors are considered, it gets still more complicated.

If the mixed, two-contributor sample had markedly different proportions of DNA, there could be a pattern of tall and short peaks across multiple loci (as in Figure 6), suggesting that the short ones represent the DNA of a "minor contributor" and the tall ones come from the DNA of a "major contributor." By using the peak heights and locations, the analyst might be able to "deconvolute" the pattern of peaks into the most likely underlying genotypes and then see if the defendant has one of these genotypes.[298] Great effort has gone into formulating a systematic procedure for determining the number of contributors to consider,[299] to listing all possible sets of genotypes that could give rise to a mixed STR profile, to scratching some off the list as inconsistent with the peak-height information, and to performing a statistical assessment of an inclusion of a defendant who has one of the genotypes left on the list. The standard algorithm for this "manual" deconvolution has six or seven steps.[300] The process requires the

297. For instance, if the three genotypes were AB/CD/AB, the A and B peaks would be twice the height of the C and D peaks, but they are not.

298. *See, e.g.,* Roberts v. United States, 916 A.2d 922, 932–35 (D.C. 2007) (holding such inferences to be admissible).

299. Decisions or assumptions about the number of contributors could be assisted by other evidence about the events that resulted in the mixture.

300. For summaries of the steps, see Jo-Anne Bright & Michael D. Coble, *Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios* 7–15 (2020); Butler et al., *supra* note 105; *see also* Frederick R. Bieber et al., *Evaluation of Forensic DNA Mixture Evidence: Protocol for Evaluation, Interpretation, and Statistical Calculations Using the Combined Probability of Inclusion*, 17 BMC Genetics 125

analyst to make mostly binary decisions about which peaks are alleles as opposed to artifacts (stutter peaks, pull-up, dye blobs).[301]

It is now generally recognized that to avoid an "expectation effect" on the analyst's judgments, the allele calls should be made before considering the genotypes of any suspects or other possible contributors to the mixed sample. Then the analyst must decide which genotypes are included or excluded from the mixture. The task becomes even more difficult when the quantity of the DNA from some contributor is marginal, with the possibility of some peaks that would be expected for larger amounts of DNA dropping out to leave only a partial profile.[302] Mixtures of degraded or inhibited DNA also are harder to interpret because parts of individual profiles may be missing. Nevertheless, it has been said that "[t]he binary method served the forensic community well for a number of years; however it was mostly restricted to single-source, two-person, and some three-person mixtures (e.g., those with a clear major component)."[303] Other observers saw manual mixture interpretation as erratic,[304] with some laboratories

(2016), https://doi.org/10.1186/s12863-016-0429-7; Peter Gill et al., *DNA Commission of the International Society of Forensic Genetics: Recommendations on the Interpretation of Mixtures*, 160 Forensic Sci. Int'l 90 (2006), https://doi.org/10.1016/j.forsclint.2006.04.009; Scientific Working Group on DNA Analysis Methods (SWGDAM), *SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories* (Jan. 12, 2017, rev. July 13, 2021), https://perma.cc/2ZCW-Q6L6.

301. On these artifacts, see *supra* section titled "Capillary electrophoresis and fluorescence."

302. On drop-out, see *supra* section titled "Did the Sample Contain Enough DNA?" For a detailed analysis of the manual interpretation steps in the context of an example with these complications, see Michael D. Coble, *Worked Mixture Example*, *in* Butler, *supra* note 113, at 537.

303. Bright & Coble, *supra* note 300, at 15.

304. An oft-cited but limited study is Itiel E. Dror & Greg Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, 51 Sci. & Just. 204 (2011), https://doi.org/10.1016/j.scijus.2011.08.004. It reports aggregated conclusions of (1) two analysts in Georgia who compared a complex four- or five-man mixture in a gang-rape case with the profiles of three suspects, and (2) 17 analysts at an unnamed laboratory given, as an abstract exercise, the mixed profile and the profile of just one of the three suspects. The two groups reached different conclusions. The extent to which the reported differences prove much about the impact of biasing information and the reproducibility of manual mixture deconvolution in general is discussed in David H. Kaye, *The Design of "The First Experimental Study Exploring DNA Interpretation,"* 52 Sci. & Just. 126 (2012), https://doi.org/10.1016/j.scijus.2011.10.003, and Itiel E. Dror, *Cognitive Forensics and Experimental Research About Bias in Forensic Casework*, 52 Sci. & Just. 128 (2012), https://doi.org/10.1016/j.scijus.2012.03.006. More systematic interlaboratory comparisons of up to five-person mixtures from 1997 through 2018 are tabulated in John M. Butler et al., *DNA Mixture Interpretation: A NIST Scientific Foundation Review* 80–81 (NISTIR 8351-DRAFT, June 2021) (tbl. 4.8), https://doi.org/10.6028/NIST.IR.8351-draft. The reasons for variability in the reports from participants in two U.S. interlaboratory studies are outlined in John M. Butler et al., *NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned*, 37 Forensic Sci. Int'l: Genetics 80, 81 (2018), https://doi.org/10.1016/j.fsigen.2018.07.024; *cf.* Michael Coble et al., *Analysis of Forensic Mixtures*, *in* Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters 49, 58–59 (Henry Ehrlich et al. eds., 2020) ("Studies from European laboratories [citations omitted] have typically shown wide variation in the results within and between

performing the task poorly.[305] Despite the degree to which "subjective" judgment can go into the determination of which peaks are real, which are artifacts, which are masked, and which are absent for some other reason, courts generally have rejected arguments that manual mixture analysis is so unreliable or so open to manipulation that the results are inadmissible.[306]

What Statistical Assessment of the Comparison to a Defendant's Profile Has Been Conducted?

When manual deconvolution of a mixed profile is complete and a defendant's profile is compatible with it, most laboratories will attempt to supply a statistical assessment to indicate how powerful the evidence is.[307] Methods for doing so fall

the participants' laboratories. [In] two interdisciplinary studies in 2005 and 2013 . . . involving laboratories performing STR genotyping in the United States and Canada . . . overall results were concordant with the studies from Europe. . . ."). Proficiency tests, which sometimes include two- or three-person mixtures, are listed in Butler et al., *DNA Mixture Interpretation: A Scientific Review*, *supra*, at 78–79 (tbl. 4.7) (reporting small error rates); *see also* Todd Bille et al., *Study of CTS DNA Proficiency Tests with Regard to DNA Mixture Interpretation: A NIST Scientific Foundation Review*, 13 Genes 2171 (2022), https://doi.org/10.3390/genes13112171 (re-analyzing the proficiency test data and concluding that "[s]ample switching, contamination, and reporting results incorrectly are serious errors [but not a consequence of] the mixture interpretation strategy").

305. For example, an inquiry into mixture interpretations at the Washington, D.C., forensic science laboratory at the behest of the U.S. Attorney's Office for the District of Columbia led to the ouster of the laboratory's management. *See supra* note 129. The common errors that were identified mostly pertained to the statistical interpretation of inclusions in mixture cases. Barber v. United States, 179 A.3d 883, 892 n.17 (D.C. 2018). Other jurisdictions have issued letters to convicted defendants notifying them that older calculations of the combined probability of inclusion (*see infra* section titled "(1) Combined Probability of Inclusion (CPI) and Exclusion (CPE)") might be inaccurate. Skinner v. State, 484 S.W.3d 434 (Tex. Crim. App. 2016); Rosado v. State, 267 So.3d 4 (Fla. Dist. Ct. App. 2019) ("The notice stated that the Broward Sheriff's Office Crime Laboratory inappropriately used Combined Probability of Inclusion (CPI) to calculate the statistical probabilities of genetic profiles in complex mixture DNA cases."); Tex. Forensic Sci. Comm'n, Summary of Texas CPI Case Review Process, Sept. 15, 2016, https://perma.cc/8MAS-EURJ.

306. United States v. Carney, No. 1:20-cr-121, 2022 WL 1155902 (S.D. Ohio Apr. 19, 2022) (reasoning that manual deconvolution for three-person mixtures satisfies *Daubert* largely because it could be tested and was in widespread use and is still in use for some mixtures in 20% of laboratories); State v. Garland, 942 N.W.2d 732 (Minn. 2020); Roberts v. United States, 916 A.2d 922, 932 n.9 (D.C. 2007) (citing cases).

307. *But see supra* note 295 (on admissibility without a statistical assessment). The SWGDAM guidelines require a quantitative statement of some kind. SWGDAM, *supra* note 300, § 3.2.1 ("Except for a reasonably assumed contributor, the laboratory shall perform statistical analysis in support of any inclusion (or a 'cannot be excluded' conclusion) irrespective of the number of alleles detected and the quantitative value of the statistical analysis.").

into two classes: inclusion–exclusion probabilities and likelihood ratios. The latter have greater support within the forensic DNA science community.[308]

Probability of inclusion or exclusion

Combined probability of inclusion (CPI) and exclusion (CPE). At first glance, it might seem that it is easy to describe the significance of a match to a mixed profile. We could simply ask what fraction of the population has genotypes that are not excluded by the alleles in the mixed profile. The answer in no way depends on the profiles of the victim, suspect, or person of interest; furthermore, no assumptions or inferences about the number of contributors are necessary for the calculation. For example, suppose we are sure that at one locus, the mixed profile indicates three alleles—A, B, and C, and no others—with population proportions 0.1, 0.2, and 0.3, respectively. The single-locus genotypes of people who could not be excluded as contributing to the profile are AA, BB, CC, AB, AC, and BC. Their (Hardy-Weinberg) probabilities are $0.1^2 + 0.2^2 + 0.3^2 + 2(0.1)(0.2) + 2(0.1)(0.3) + 2(0.2)(0.3) = 0.36$.[309] So 36% of the population could not be excluded at this locus. Performing the same calculation at the other loci and multiplying them (linkage equilibrium) gives the "combined probability of inclusion,"[310] often referred to as CPI or RMNE (for "random man not excluded," although its use is not restricted to male contributors). The combined probability of exclusion (CPE) is 1 minus the CPI.

Despite its computational simplicity,[311] CPI has been deprecated as "hard to justify"[312] and an "interim" measure.[313] In the three-allele mixture, for example,

308. *See infra* section titled "Likelihood ratios (LRs)."

309. *See supra* note 169.

310. Adjustments for population structure (see *supra* section titled "Could an Unrelated Person Be the Source?") can be applied. Examples of how to compute CPI in different scenarios are given in SWGDAM, *supra* note 300, § 4C.

311. Determining the set of loci and alleles to use in the computation can be anything but simple. On the subtleties in using CPI properly, see *Barber v. United States*, 179 A.3d 883, 892 n.17 (D.C. 2018); Bieber et al., *supra* note 300; Coble et al., *supra* note 264, at 243–44; SWGDAM, *supra* note 300, § 4C.

312. NRC II, *supra* note 7, at 130.

313. Bieber et al., *supra* note 300, at 3. PCAST, *supra* note 105, at 82, expressed skepticism of the "foundational validity" of CPI, and stressed that "DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping." The group added that "[i]f, for a limited time, courts choose to admit results based on the application of CPI, validity as applied would require that, at a minimum, they be consistent with the rules specified in the paper."

it counts pairs of people who could not jointly be the two contributors.[314] Thus, it does not actually resolve the mixture profile into the genotypes of those individuals who probably contributed to the mixed sample. Moreover, it disregards peak-height information, and it ignores the fact that there may be a known contributor such as the victim.

The number of reported opinions responding to challenges to CPI as lacking scientific validity or general acceptance is relatively small. Some courts have reasoned that the CPI is no different than the random-match probability accepted in single-source cases.[315] More often, courts recognize that the CPI has been criticized in scientific literature but conclude that it remains generally accepted as a reasonable or often conservative way to express the significance of an inclusion.[316]

Random-match probability (RMP). Specifying the number of contributors in the mixture allows the expert to compute a somewhat different probability of inclusion, called the random-match probability (RMP). The RMP also can use peak heights and other information to deconvolve the mixed profile in light of any known or assumed contributor profiles. After taking all these things into account, the analyst will be left with a reduced set of plausible genotypes of the contributors. The RMP is just the sum of the probabilities for each of these genotypes. For example, suppose the victim's account and other information in a rape case establishes that one man committed the rape, and the victim's genotype at a locus is AB. DNA from a vaginal swab demonstrates the presence of three alleles A, B, and C with the peak for A and B being about the same height and C being much lower. The male contributor's genotype could be AC, BC, or CC,[317]

314. For instance, the genotype pairs AA/AA, AA/BB, AA/CC, AA/AB, and AA/AC can be ruled out. They can produce no more than two peaks in the mixture electropherogram.

315. *E.g.*, People v. Sandifer, 65 N.E.3d 969, 980 (Ill. App. Ct. 2016); *cf.* Coy v. Renico, 414 F. Supp. 2d 744, 762 (E.D. Mich. 2006) (as "mere[] extensions of the generally accepted and generally reliable techniques applied to single source DNA samples," CPI and CPE satisfy due process).

316. *E.g.*, State v. Bander, 208 P.3d 1242 (Wash. Ct. App. 2009); *cf.* State v. Garland, 942 N.W.2d 732, 744 (Minn. 2020) (relying on unspecified "evidence before the district court . . . including a peer-reviewed scientific article regarding best practices for using CPE" to conclude that "CPE has long been accepted by the forensic community as a valid method for calculating the probability of exclusion from a DNA mixture").

317. The rapist must be responsible for the small peak C. Because it is small, the man is a minor contributor to the sample. A genotype of CC would explain the pattern of major peaks for A and B and a minor peak for C. Genotypes AC and BC for the rapist also would explain this pattern. For a small quantity of DNA of type AC, the A peak gets boosted slightly but remains approximately equal to the B peak. For a minor contribution of genotype BC, the B peak gets boosted, but only slightly. Because no other male genotypes (AA, AB, and BB) can explain the minor peak C in the mixture profile, the rapist must be genotype AC, BC, or CC, as stated in the text.

and the expert could estimate the frequency of these genotypes (using the population allele frequencies posited in the previous subsection) as $2(0.1)(0.3) + 2(0.2)(0.3) + (0.3)(0.3) = 0.27$. Instead of reporting the 36% for the single-locus inclusion probability as in the CPI calculation, the expert could report a smaller inclusion probability of 27%.[318]

Testimony about the RMP (or the frequency with which unrelated individuals would match the deduced, smaller set of genotypes) is often encountered.[319] Sometimes the RMP or frequency is described as if it were the CPI.[320] Issues likely to arise with the RMP are whether identification of major and minor alleles is correct and how the number of contributors was determined.

Likelihood ratio (LR)

Definition of the likelihood ratio. Rather than testify to inclusion probabilities for close relatives or for randomly selected, unrelated members of the suspect population, a DNA expert can present a likelihood ratio (LR) to express the evidentiary value of a DNA comparison.[321] In the most general formulation,

318. More examples and more refined calculations are in SWGDAM, *supra* note 300, § 4A; *see also* Bright & Coble, *supra* note 300, at 12–14. In two-person mixture cases in which it is reasonable to assume that the victim's DNA is present, the other contributor's genotype might be clear from the remaining peaks, and the RMP will be the same as in a single-source case for that contributor.

319. *E.g.*, United States v. Carney, No. 1:20-cr-121, 2022 WL 1155920 (S.D. Ohio Apr. 19, 2022) (RMP for the major contributor to a three-person mixture was said to be "1 in 299 septillion 400 sextillion" and therefore dispositive even though no alleles from the minor contributors could be reported); United States v. Graves, 465 F. Supp. 2d 450, 453 (E.D. Pa. 2006) (RMPs of 1/2, 1/2,900, and 1/3,600); People v. Roberts, 283 Cal. Rptr. 3d 357 (Cal. Ct. App. 2021); State v. Gonzalez, 204 A.3d 1183, 1189 (Conn. App. Ct. 2019); State v. Lowery, 427 P.3d 865 (Kan. 2018); State v. Hannon, 703 N.W.2d 498, 508–09 (Minn. 2005).

320. *E.g.*, People v. Brown, 82 N.E.3d 148, 169–70 (Ill. App. Ct. 2017) (The state's expert "testified that based on the only 6-loci analysis, he could determine only that defendant could not be excluded as a contributor to the major profile. . . . [T]he frequency which estimated the chance a random person would be included as a contributor in that major DNA profile [was] 1 in 670,000 black, 1 in 580,000 white, or 1 in 6.1 million Hispanic unrelated individuals. . . . The DNA Advisory Board has endorsed two methods . . . in cases of mixed DNA samples: (1) the combined probability of inclusion (or its reverse, the combined probability of exclusion) or (2) the likelihood ratio calculation."). It is not always clear from an opinion whether the expert is giving a CPI or an RMP. *E.g.*, State v. Dwyer, 985 A.2d 469, 478–79 (Me. 2009) (giving "inclusion probabilities" without indicating whether they were CPIs or RMPs). As a result, some of the opinions cited above as illustrative of one method or the other could be misclassified.

321. In civil and criminal cases that involve DNA testing for parentage, LRs are routinely introduced under the name "paternity index." 1 McCormick on Evidence, *supra* note 1, § 211.

an LR is the ratio between (1) the probability of the observed data when one hypothesis about the origin of the data is true and (2) the probability of the data when a non-overlapping hypothesis is true.[322] In symbols,

$$\mathrm{LR} = \frac{\mathrm{P}(\mathrm{data} \mid \mathrm{H}_1)}{\mathrm{P}(\mathrm{data} \mid \mathrm{H}_2)}$$

where the vertical line stands for "given" or "conditional on." For DNA evidence, H_1 could be a "contributor hypothesis" that names the defendant as the contributor or one of the contributors, and H_2 could be a "noncontributor hypothesis" that names individuals other than the defendant as the contributors. When multiple contributors are likely, various pairs of hypotheses may need to be considered to avoid misstating the value of the evidence.[323]

The baseline for understanding the value of the LR is 1—the ratio when the probabilities are equal. Equal evidence probabilities mean that the evidence is not probative of which hypothesis is true. For example, if H_1 asserts that a defendant's DNA is in a trace-evidence sample, and H_2 asserts that the defendant's identical twin's DNA is in the sample, then LR = 1. The DNA data have the same probability under each scenario and therefore are not probative of which twin is a contributor to the sample.

The value of the LR normally would be different if the noncontributor hypothesis H_2 were that a different relative or an unrelated person is a contributor. Consider a single-source sample from the crime scene: If the STR profile of the sample is more probable if the defendant's DNA is in it than if the alternative person's DNA is there, then the LR exceeds 1. The profile data are more compatible with the defendant's being the source (H_1) than with the alternative (H_2). As such, the data can be said to support H_1 (compared to H_2). Conversely, if the LR is less than 1, the data support H_2.

In the legal literature, LRs often are presented as a way to quantify probative value.[324] The idea is that relevant evidence changes the probability of some material fact, and the probative value of this evidence is the extent of the change. As noted in "Appendix: Conditional Probability and Bayes' Rule" of the *Reference Guide on Statistics and Research Methods*, in this manual, under certain conditions, the change in the odds (known to statisticians as the Bayes' factor) is simply the LR. When this is the case, the odds after receiving the evidence (the posterior odds) are just the odds before receiving the evidence times the LR:

$$\text{posterior odds} = \mathrm{LR} \times \text{prior odds}$$

322. Mathematically, when the variable is essentially continuous (such as the height of a peak on an electropherogram) rather than discrete (such as the number of tandem sequences in an STR allele deduced from a peak), the LR is a ratio of probability densities. The definition above also omits a proportionality constant in both the numerator and denominator that cancels out when the ratio is formed.

323. Richard Wivell et al., *An Investigation into Compound Likelihood Ratios for Forensic DNA Mixtures*, 14 Genes 714 (2023), https://doi.org/10.3390/genes14030714.

324. *See* 1 McCormick on Evidence, *supra* note 1, § 185.

This version of Bayes' rule reinforces the idea that evidence with an LR of 1 has no probative value. Multiplying by 1 has no effect on the odds of H_1 relative to H_2.

LRs are versatile. With DNA evidence, they can be used to convey the probative value of an inclusion for complex, mixed samples as well as for single-source samples with respect to a variety of alternative explanations for the composition of the sample. For instance, the hypothesis advanced by the prosecution might be that the "defendant and his deceased wife . . . were contributors to a DNA mixture profile drawn from a blood stain on defendant's sweatshirt," and the contrasting hypothesis could be that DNA from "two randomly selected individuals" was in the stain.[325]

Moreover, LRs do not require a binary, include-exclude determination. They do not even require analytical and stochastic thresholds for allele determinations. As such, they can reflect more of the information in the electropherograms than both the CPI and the RMP can. Many forensic geneticists and statisticians regard them as particularly advantageous for mixture interpretation,[326] and laboratories across the world have turned to probabilistic-genotyping software to produce LRs to interpret electropherograms from low-template, degraded, or mixed DNA samples with greater consistency than unassisted human judgment can provide.[327]

In evaluating this development, it is important to keep in mind that the LR is merely a mathematical expression. The utility of any stated value for the quantity depends on how it was computed. Are the hypotheses or propositions appropriate to the case at hand? This is a question of relevance. Has the computational system been shown to give numbers that properly discriminate between the propositions for the DNA samples in the case? This is a question of validity. Assuming validity, how should LRs be presented to lay factfinders so that they

325. People v. Ramsaran, 79 N.E.3d 1120 (N.Y. App. Div. 2017).

326. *E.g.*, Bright & Coble, *supra* note 300, at 11 ("The LR is our preferred approach to interpreting forensic DNA mixtures."); Royal Soc'y, Forensic DNA Analysis: A Primer for Courts 36 (2017) ("Likelihood ratios are generally accepted as being the most appropriate method for evaluating the evidential strength of DNA profiles."); Peter Gill et al., *DNA Commission of the International Society for Forensic Genetics: Assessing the Value of Forensic Biological Evidence—Guidelines Highlighting the Importance of Propositions: Part I: Evaluation of DNA Profiling Comparisons Given (Sub-) Source Propositions*, 36 Forensic Sci. Int'l: Genetics 189 (2018), https://doi.org/10.1016/j.fsigen.2018.07.003; Peter Gill et al., *DNA Commission of the International Society of Forensic Genetics: Recommendations on the Evaluation of STR Typing Results that May Include Drop-out and/or Drop-in Using Probabilistic Methods*, 6 Forensic Sci. Int'l: Genetics 679, 684 (2012), https://doi.org/10.1016/j.fsigen.2012.06.002 ("probabilistic approaches and likelihood ratio principles are superior to classical methods"). United Kingdom guidelines regard statements of LRs as the only acceptable approach to mixture interpretation. Forensic Science Regulator, Guidance: DNA Mixture Interpretation (FSR-G-222 Issue 3) (2020), https://perma.cc/L34N-Y54X.

327. For a concise history of the transition to probabilistic genotyping systems, see Michael D. Coble & Jo-Anne Bright, *Probabilistic Genotyping Software: An Overview*, 38 Forensic Sci. Int'l: Genetics, 219, 219–21 (2019), https://doi.org/10.1016/j.fsigen.2018.11.009.

fairly convey the weight of the evidence with respect to the relevant hypotheses? This is a question of probative value and prejudice. The remaining subsections elaborate on these points and describe the ways in which LRs for DNA evidence are computed.

Binary genotyping. Traditional forensic STR-CE profiling is "binary" or "deterministic." Like a baseball umpire calling a pitch a strike or a ball, an analyst (or software) labels a peak as either an allele or not. An allele call requires that the peak height exceed an analytical threshold and that the peak not look like an artifact.[328] At each locus, the analyst decides whether there is one allele (homozygosity) or two different alleles (heterozygosity) on the pair of paternal and maternal chromosomes. If the reported profiles in two samples are consistent, there is a profile match. At that point, taking the reported matching profile as 100% certain, analysts can quote LRs for inclusions as the reciprocal of the RMP.[329] Thus, the data from the laboratory instruments are processed in two phases—a binary match/no-match phase, and a separate phase to give an LR for the categorical match (by regarding the analyst's inclusion of the defendant as an adequate summary of the data).

This two-stage procedure works well enough in many cases (and in the absence of better alternatives). The 1996 NRC committee report endorsed then-available binary LRs (based solely on the positions of DNA fragments subjected to electrophoresis) as opposed to the CPI.[330] SWGDAM promulgated guidelines for calculating these LRs.[331] Despite objections based on *Frye* or *Daubert*, appellate courts that have considered binary LRs in mixture cases have upheld their admission.[332]

328. *See supra* section titled "Miniaturization for rapid capillary electrophoresis" n.52 (before Table 5); *supra* section titled "Did the Sample Contain Enough DNA?"

329. *E.g.*, Augillard v. Madura, 257 S.W.3d 494, 498–99 (Tex. Ct. App. 2008) ("the test showed a complete match at all seventeen DNA markers with a likelihood ratio exceeding one trillion"). The LR is the reciprocal of the RMP if the match, as determined by thresholds and human judgment, has probability 1 when H_1 is true (if the defendant is a source, the analyst is certain to report an inclusion) and probability RMP when H_2 is true (if a random, unrelated person is a source, the probability of inclusion is just the random-match probability).

330. NRC II, *supra* note 7, at 129 ("when the contributors to a mixture are not known or cannot otherwise be distinguished, a likelihood-ratio approach offers a clear advantage and is particularly suitable").

331. SWGDAM, *supra* note 300, § 4(B).

332. *E.g.,* State v. Garcia, 3 P.3d 999 (Ariz. Ct. App. 1999) (applying *Frye*); Commonwealth v. Gaynor, 820 N.E.2d 233, 252 (Mass. 2005); People v. Coy, 669 N.W.2d 831, 835–39 (Mich. Ct. App. 2003) (incorrectly treating mixed-sample likelihood ratios as a part of the statistics on single-source DNA matches that had already been held to be generally accepted); *cf.* Coy v. Renico, 414 F. Supp. 2d 744, 762–63 (E.D. Mich. 2006) (likelihood ratio for a mixed stain in *People v. Coy*, *supra*, was consistent with due process).

Probabilistic genotyping software (PGS). Binary LRs have significant limitations. Using only the genotypes as reported from the initial match/no-match phase overlooks any uncertainty in the allele calls and related single-locus genotype determinations. In particular, it ignores the possibility of allelic drop-out or drop-in, which could make the assignment of possible genotypes less than certain. It tosses out any alleles that might be producing peak heights below the laboratory's analytical threshold (or that are wrongly thought to be artifacts).[333] In effect, a peak above the line is treated as if it has probability 1 for being an allele, while a peak just below the line is treated as if it has probability zero of being an allele. But surely, if an allele is present (or absent) in a sample, then the probability of a peak just below the threshold is not radically different from a peak just above the line. Newer, probabilistic genotyping systems overcome these limitations by computing a likelihood ratio $P(\text{data}|H_1) \div P(\text{data}|H_2)$ without having to declare a match or inclusion.

They do so by using weighted averages, for each hypothesis, of the probabilities that all sets of genotypes that could be in the complicated mixture profile are in fact there—based on how often the possible genotypes occur in the relevant population.[334] These "prior probabilities" are weighted by the probability that the genotype set (if present) would give rise to the pattern in the electropherogram.[335] Thus, a possible genotype set that is more likely to give rise to the electropherogram if it is assumed to be in the mixture profile gets more weight.[336]

Over the past two decades, researchers have developed a variety of PGS programs with increasingly complete statistical models of the production of true peaks and artifacts in electropherograms. The simplest programs (called qualitative, discrete, or semi-continuous) consider only the alleles that might be

333. *See, e.g.,* Butler, *supra* note 113, at 39–40.

334. For example, under a contributor hypothesis H_1, such as "the defendant and a brother are the source of the semen," there might be just one set of genotypes at a given locus that conceivably could have produced the electropherogram. For concreteness, let's say this set is AB/AC, meaning that one brother is AB and the other is AC. Under a noncontributor hypothesis H_2 of two random, unrelated men, there might be two sets, such as AB/AC and AA/BC. The prior probability of each of these genotype sets comes from population-genetic models and allele-frequency databases.

335. The genotypes-to-data probability, P(data|genotypes), depends entirely on the sample and the CE measurement process. It is the same under the contributor and the noncontributor hypotheses.

336. The mathematical expression for the weighted averaging is in Peter Gill et al., *A Review of Probabilistic Genotyping Systems: EuroForMix, DNAStatistX and STRmix™*, 12 Genes 1559, at 2 (2021), https://doi.org/10.3390/genes12101559, and Mark W. Perlin & Alexander Sinelnikov, *An Information Gap in DNA Evidence Interpretation*, 4 PLoS ONE e8327, at 2 (2009), https://perma.cc/RZ3T-KLHV.

present[337] and do not model the height of the peaks.[338] They incorporate estimated probabilities of allelic drop-in and drop-out[339] when assigning weights on the prior probabilities.[340] The output is an LR that draws on more of the information in the mixture electropherogram.[341] One such program, the Forensic Statistical Tool developed at the New York City Office of the Chief Medical Examiner, was the subject of considerable litigation[342] and is no longer used.[343]

These programs have been supplanted by "continuous" or "quantitative" systems that model peak heights themselves, automatically accounting for drop-in and drop-out as well as stutter peaks, degradation, and other complicating factors. Two types of continuous systems have been developed. One type posits an explicit mathematical function relating to peak heights. It uses a common statistical method for estimating parameters and applies the estimates to find the genotypes-to-data probabilities.[344] The other type relies on Bayes' rule to obtain these weights.[345]

337. Even in this regard, many "semi-continuous methods do not model artefacts such as stutter. In these models, peaks must be assigned as stutter or labelled as allelic by an analyst prior to interpretation [by software]." Jo-Anne Bright et al., *A Series of Recommended Tests When Validating Probabilistic DNA Profile Interpretation Software*, 14 Forensic Sci. Int'l: Genetics 125, 126 (2015), https://doi.org/10.1016/j.fsigen.2014.09.019.

338. At most, they use peak heights in more limited ways, "to inform the probability of drop-out or to infer a major donor genotype by applying different drop-out probabilities per contributor." Nevertheless, "[t]hese systems do represent an advance over the binary model as they can take account of multiple contributors, low-template DNA and replicated samples." Gill et al., *supra* note 336, at 3.

339. Drop-in and drop-out probabilities come from experiments on how often peaks from samples with known alleles disappear (drop below the analytical threshold) or appear when the known samples do not contain known alleles for those peaks. *See* Coble & Bright, *supra* note 327, at 221.

340. For a numerical example of how drop-in and drop-out probabilities are incorporated, see Michael Coble et al., *Mixtures and Probabilistic Genotyping*, *in* Forensic DNA Applications: An Interdisciplinary Perspective 155, 164–66 (Dragan Primorac & Moses Schanfield eds., 2d ed. 2023).

341. For details on the operation of a leading qualitative system, see Gill et al., *supra* note 36, at 129–80 (on LRmix and LRmix Studio); *see also* Bright & Coble, *supra* note 300, at 107–40.

342. *Compare*, *e.g.*, United States v. Jones, 965 F.3d 149, 162 (2d Cir. 2020) ("no abuse of discretion in the district court's conclusion [of] reliability sufficient to support admission" of an LR likelihood ratio of "1,340, showing very strong support"), *with* State v. Rochat, 269 A.3d 1177 (N.J. App. Div. 2022) (FST not generally accepted).

343. *Jones*, 965 F.2d at 152.

344. On the general nature of parameters in a statistical model and their estimation, see David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual. In one widely used program, EuroForMix, the statistical model for differences between estimated and expected peak heights is known as the gamma (γ) distribution. Gill et al., *supra* note 36, at 199–200; Gill et al., *supra* note 336, at 13–14. The parameters of this distribution are closely related to the proportion of mixture DNA from each contributor and the coefficient of variation of peak heights across all the loci. Gill et al., *supra* note 36, at 201–05. The parameter values are estimated by the maximum-likelihood method. *Id.* at 461.

345. Gill et al., *supra* note 36, at 462. Because peak height is a continuous variable, the version of Bayes' rule for discrete variables presented in the *Reference Guide on Statistics and Research Methods*,

The two programs that are most commonly used in the United States are of the latter type. They have the brand names TrueAllele and STRmix. To approximate the genotypes-to-data probabilities, they take a kind of random walk through the space of all possible parameter values,[346] spending more time at those combinations at which the pattern of peaks computed by the model come closest to the observed values.[347] The relative time spent at each such combination determines the weights P(data|genotype set).[348] The technical name for this probabilistic approximation method is Markov chain Monte Carlo (MCMC).[349] The "Monte Carlo" part refers to computer simulation of the outcomes of any process that has some randomness in it.[350] Simulations rarely give exactly

in this manual, needs to be restated with a probability density function and integration instead of summation. The underlying concepts are the same, however, and, for ease of presentation, this reference guide has not distinguished between discrete probability distributions and probability density functions. The LR for continuous models is a ratio of probability densities.

346. For partial descriptions of the parameters in the peak-height model of STRmix, see Bright & Coble, *supra* note 300, at 143 ("For the simplest profile . . . these are 1. DNA amount (template) for each contributor to the profile 2. The degradation rate for each contributor to the profile 3. Amplification efficiencies for each locus within the profile."); Coble et al., *supra* note 340, at 167 ("template quantity, degradation, stutter ratios, etc.").

347. At least, that is the expectation. The game of "hot and cold," in which a person searching for an object in a room is told whether he is getting closer to or farther from the object as he moves around the room is sometimes offered as a metaphor for the computer-simulation procedure. *E.g.*, United States v. Gissantaner, 417 F. Supp. 3d 857 (W.D. Mich. 2019), *rev'd*, 990 F.3d 457 (6th Cir. 2021); United States v. Lewis, 442 F. Supp. 3d 1122, 1142 (D. Minn. 2020) (Rep. & Recommendation) (extending the analogy to multiple rooms). A better (but still imperfect) analogy might be trying to climb the highest mountain in a region with no visibility by taking proposed steps that increase the elevation with a higher probability than steps that do not, until no proposed step increases the altitude. This should get the climber to the summit if the mountain is smooth and if there are no smaller mountains in the region. Unlike the hot-and-cold game in which one *knows* when the object is found, in the latter landscape, there is a risk of being fooled into thinking that a smaller peak is the highest there is.

348. *See* Coble et al., *supra* note 340, at 167–68.

349. The "Markov chain" part refers to situations in which the next value of a random variable depends to some degree on just the previous value (like a peculiar coin that is a little more likely to turn up heads when tossed if a head had turned up on the previous toss). "Since their invention [in 1905], Markov chains have become extremely important in a huge number of fields such as biology, game theory, finance, machine learning, and statistical physics." Joseph K. Blitzstein & Jessica Hwang, Introduction to Probability 459 (2015). Nonetheless, the success of an MCMC algorithm in one domain or type of problem does not guarantee that it will work as well in another situation.

350. For example, if we wanted to estimate the probability of a lucky streak of ten heads when tossing a fair coin, we could program a computer to pick a 0 or a 1 at random ten times and see if the sequence was 1111111111. One such simulation would not tell us much, but if the computer simulated ten tosses one million times, keeping track of how many trial sequences were 1111111111, the proportion usually would be close to the true value of the probability. That the inventors of the computer-simulation method referred to a famous gaming casino in naming it is an historical accident. *See* Nicholas Metropolis, *The Beginning of the Monte Carlo Method*, Los Alamos Sci., Special Issue 125, 127 (1987), available at https://perma.cc/M53S-Y9X9.

identical estimates if they are rerun, and "[a] simulation result is easy to criticize: how do you know you ran it long enough? How do you know your result is close to the truth? How close is 'close?'"[351]

These are questions to be addressed in validation studies, considered in the next subsection. Proof of validity also needs to attend to the more fundamental issue of the adequacy of the parametric models for predicting the observed data from the possible genotype sets. PGS programs that use Bayesian methods and MCMC computations may not use exactly the same statistical models, and one program may require more preliminary input from the DNA analyst or laboratory using it than another.[352] Rather than a blind acceptance of the output, the reasonableness and impact of these choices may need to be examined.

Validation of PGS programs. In the courtroom, the most important output of PGS is the LR for a set of hypotheses that include the defendant as a contributor to the mixture (or to a single-source sample, typically of very low quantity).[353] The LR is not a measured physical or chemical property, like height or weight. Rather, it is an expression of the degree to which the electropherogram is more indicative of scenarios that include the defendant as a contributor than it is of scenarios that do not. How, then, do scientists who have written the software convince other scientists that the values of the LR are believable?

Validation has at least three components. First, the mathematical equations used in the software and the ways of solving those equations must be appropriate.[354]

351. Blitzstein & Hwang, *supra* note 349, at 421 (also noting that "[t]he simulations may need to run for a vast, unknown amount of time to get decent answers").

352. *See* Daniels v. State, 312 So.3d 926 (Fla. Dist. Ct. App. 2021) (whereas STRmix requires laboratory-specific values from "internal validation" studies, TrueAllele does not). TrueAllele also dispenses with analytic thresholds; but STRmix uses them. William C. Thompson, *Uncertainty in Probabilistic Genotyping of Low Template DNA: A Case Study Comparing STRMix™ and TrueAllele™*, J. Forensic Sci. (2023), https://doi.org/10.1111/1556-4029.15225. However, STRmix may be moving away from these rigid cutoffs. *See* Duncan Taylor & John Buckleton, *Combining Artificial Neural Network Classification with Fully Continuous Probabilistic Genotyping to Remove the Need for an Analytical Threshold and Electropherogram Reading*, 62 Forensic Sci. Int'l: Genetics 102787 (2023), https://doi.org/10.1016/j.fsigen.2022.102787.

353. Some PGS programs are also useful for estimating the number of contributors and mixture ratios and for producing a list of high-LR contributor genotypes that could be submitted to an offender DNA database.

354. The conceptual underpinnings of the two main continuous PGS systems have not attracted major criticism. The weighted averaging described in the section titled "Probabilistic genotyping software (PGS)" comes out of the basic mathematics of probability theory and Bayes' rule. The values for the prior probabilities are population genotype frequencies that are calculated from population-genetics models and are discussed in the section titled "Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing" above. The modeling of the genotypes-to-data probabilities also

In litigation, the general mathematical structure behind most PGS software has not been seriously questioned.[355]

Second, the coding of the mathematical formulas and techniques for solving the equations must be correct. Examining parts of the source code of a program and testing whether modules work as they are supposed to as the program is written and revised can reveal coding and logical problems.[356] Thus, voluntary quality-assurance standards for software (and hardware) "verification and validation" are well known among software engineers.[357] These standards are sometimes cited to support motions for discovery of all of the source code of a PGS program or to buttress an argument that the program cannot be found to be scientifically valid or generally accepted without publishing the code (open-access software), or at least affording defense experts extensive review of proprietary software, perhaps under a confidentiality order.[358]

needs to be sound. Finally, the statistical procedures (like an MCMC algorithm, which is not itself the genetic or statistical model at the heart of PGS) must be capable of solving (at least approximately) the equations to which they are applied.

355. *E.g.*, United States v. Lewis, 442 F. Supp. 3d 1122, 1145 (D. Minn. 2020) (Rep. & Recommendation) ("The underlying principles on which STRmix is based—the MCMC, the Hastings-Metropolis algorithm, and Bayesian statistics—are not themselves challenged as unreliable. Nor is the purely mathematic calculation of relative probabilities (i.e., the likelihood ratio)."); People v. Davis, 290 Cal. Rptr. 3d 661, 679 (Cal. Ct. App. 2022) ("[t]he scientific and mathematical principles behind STRmix are well-established and widely-accepted in the scientific community"); People v. Wakefield, 195 N.E.3d 19, 28 (N.Y. 2022) ("It was undisputed that the foundational mathematical principles [of TrueAllele] (MCMC and Bayes' theorem) are widely accepted in the scientific community."). *But see* United States v. Gissantaner, 417 F. Supp. 3d 857, 870 (W.D. Mich. 2019), *rev'd*, 990 F.3d 457 (6th Cir. 2021) (suggesting that software that implements Bayes' rule is problematic because "Bayesian reasoning 'breaks down in situations where information must be conveyed from one person to another such as in courtroom testimony'").

356. "Black box" testing, which involves no inspection of the underlying code, is also a standard procedure. *See, e.g.*, Christopher Henard et al., *Comparing White-Box and Black-Box Test Prioritization*, ICSE '16: Proceedings of the 38th International Conference on Software Engineering 523 (2016), https://doi.org/10.1145/2884781.2884791.

357. In this context, "verification" refers to making sure that the software does what the specifications call for. "Validation" refers to ensuring that the product meets the needs of its users. The ISO/IEC/IEEE 29119 series of standards offer guidance on "V & V" for business and other environments. *See* ISO/IEC/IEEE 29119–1:2022(en) (Software and systems engineering—Software testing—Part 1: General concepts). They go beyond establishing scientific or technical merit.

358. *See, e.g.,* People v. Wakefield, 195 N.E.3d 19 (N.Y. 2022); State v. Watkins, 648 S.W.3d 235 (Tenn. Ct. Crim. App. 2021); *cf.* Nathaniel Adams et al., *Letter to the Editor—Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems*, 63 J. Forensic Sci. 339 (2018), https://doi.org/10.1111/1556-4029.13687 (urging developers to utilize the IEEE V & V standards). Without necessarily relying on the V & V software engineering standards, legal academics tend to advocate disclosure of source code. *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343, 1376 (2018) (citing articles). The authors of some PGS programs are more skeptical of the need for defendants' experts or third parties to peruse the code itself. *E.g.*, Duncan A. Taylor et al., *Commentary: A "Source" of*

The third component of validation is direct testing of the PGS by presenting "a wide range of evidence types, representing a typical case, as well as extreme examples."[359] The performance on data from a large number of mixtures of known composition can provide convincing empirical proof of the validity of a conceptually sound system (for mixtures like those in the experiment). Mixed samples can be constructed in the laboratory by combining DNA in varying ratios and total amounts from known individuals.[360] Mixed profiles for validation also can be simulated by combining known single-sample-profile data together mathematically.[361] Yet a third approach uses samples from cases in which defendants were convicted.[362] The idea in each approach is to assemble two sets of paired data for testing the PGS: contributor-mixture pairs and

Error: Computer Code, Criminal Defendants, and the Constitution, 8 Frontiers Genetics 33 (2017), https://doi.org/10.3389/fgene.2017.00033 (maintaining that it is "nearly impossible to identify subtle errors in code by viewing the code").

A different argument about PGS source code is that studying it is somehow like cross-examining an "artificial intelligence," and the Confrontation Clause therefore gives the defendant in a criminal case the right to see the code itself when the AI's output is testimonial under *Crawford v. Washington*, 541 U.S. 36 (2004). *Wakefield*, 195 N.E.3d at 44–45 (concurring opinion). A more straightforward constitutional analysis would look to the Due Process Clause and ask whether the need to inspect the source code is critical in light of other methods of testing the operation of the software.

359. Gill et al., *supra* note 336, at 10.

360. Samples also can be made from DNA with different degrees of degradation or other complicating factors.

361. Artificially generating the validation profiles rather than profiling "made-up samples" has been advocated to "control the input variables exactly and produce profiles where the expected answer is known. [I]t is effectively impossible to create an exact 1:1 [or other mixture-ratio] mixture in vitro, [whereas artificially mixed profiles] are simply tidy single source, major, minor and balanced profiles." Bright et al., *supra* note 337, at 126.

362. Treating these defendants as true contributors gives us a set of presumed-contributor-mixture pairs for testing the PGS. A much larger set of known noncontributor-mixture pairs for testing can be created by substituting genotypes of the defendants in all the other (unrelated) cases for the actual defendant. For a study employing this design, see Mark W. Perlin et al., *New York State TrueAllele® Casework Validation Study*, 58 J. Forensic Sci. 1458, 1469 (2013), https://doi.org/10.1111/1556-4029.12223; *cf.* David H. Kaye et al., *Validating the Probability of Paternity*, 31 Transfusion 823 (1991), https://doi.org/10.1046/j.1537-2995.1991.31992094670.x (examining the empirical distribution of the likelihood ratio for HLA-typing results in civil paternity cases). One might object that the presumed-contributor-mixture pairs could include some falsely convicted defendants. Obviously, it would be better to restrict the true-pair group to only true (contributor-mixture) pairs, as can be done with manufactured mixtures. However, a fraction of noncontributor-mixture pairs in the presumed-contributor-mixture pairs can only make it harder to find that a valid PGS typically generates large LRs for the group designated as true pairs. A clear difference in PGS performance, as between the two large groups of pairs, is thus still good evidence of validity. A failure to find a difference is more ambiguous. It could be explained as either a lack of validity or as a consequence of having too many falsely convicted offenders' profiles in the true-pairs group.

noncontributor-mixture pairs.[363] Each method of forming the two sets has advantages and disadvantages, and a series of studies with the different designs is best.

The PGS can be applied to these validation test sets in several ways. Analysts can verify that the program gives large LRs for true pairs in simple cases (ones they can readily resolve without the software).[364] As the information available from the mixture declines (because of more complex peaks and low-template samples), the LR values should drop as well. Beyond such rudimentary checks, rates of misleading LRs can be determined.[365] The observed proportion of cases in the validation test set for which the PGS LR points in the wrong direction (toward exclusion for contributors or toward inclusion for noncontributors)[366] is a kind of "error rate" for the PGS.[367]

Nonetheless, these misleading-LR rates do not tell the whole story. Even if the overall rates of misleading evidence are small, how do we know that an LR of 1 million is dramatically more powerful than an LR of, say, 1,000 or 100?[368] To address this question, validation studies can examine whether larger LRs are indeed associated with smaller rates of misleading evidence. Do LRs of 1,000 or

363. The false-pair cases can be formed by pairing the data for each mixture with many randomly generated genotypes (using the allele frequencies in any population of interest).

364. Bright et al., *supra* note 337, at 126.

365. *See* Richard Royall, *On the Probability of Observing Misleading Statistical Evidence*, 95 J. Am. Stat. Ass'n 760 (2000).

366. The tendency of the computed LRs to exceed 1 for mixtures that include the defendant's DNA has been called "sensitivity." The tendency to be less than 1 for mixtures that do include the defendant has been called "specificity." *E.g.*, Scientific Working Group on DNA Analysis Methods (SWGDAM), Guidelines for the Validation of Probabilistic Genotyping Systems § 21 (2015), accessible via https://perma.cc/5V7F-6VJ7. The terms are an extension of their more established meaning for binary tests. *See supra* section titled "Validity."

367. United States v. Gissantaner, 990 F.3d 457, 465 (6th Cir. 2021); Colin Aitken & Franco Taroni, *The History of Forensic Inference and Statistics: A Thematic Perspective, in* Handbook of Forensic Statistics 3, 23 (David Banks et al. eds., 2021); David H. Kaye, *Forensic Statistics in the Courtroom*, *in id.* at 225. Other measures of performance also are in use. *See* Aitken & Taroni, *supra*, at 23–24; Daniel Ramos et al., *Validation of Forensic Automatic Likelihood Methods*, *in* Handbook of Forensic Statistics 143 (David Banks et al. eds., 2021). Repeatability of computed LRs (also referred to as "precision" and "reproducibility" in the standards and writing on PGS) can be studied by rerunning the PGS on the same data. It is an issue for MCMC computations. *See, e.g.*, David W. Bauer et al., *Validating TrueAllele Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors*, 65 J. Forensic Sci. 380 (2020), https://doi.org/10.1111/1556-4029.14204.

368. The court in *United States v. Lewis*, 442 F. Supp. 3d 1122 (D. Minn. 2020), suggested that proof that "[t]he variation in the precise LR numbers generated by STRmix from one run to the next is very small" (deviating by no more than a factor of ten) answered this question. *Id.* at 1130. But replicability must not be mistaken for validity. Generating similar results on repeated tries is not necessarily the same as generating correct results. A scale that always reported objects with true weights of 1 gram and 1 kilogram as weighing 1–10 grams (first object) and 100–1,000 kilograms (second object) would massively overstate the weight of the heavier object.

more crop up in the noncontributor-mixture cases more than one time in 1,000? Do LRs of 1 million or more arise more often than one time in a million in these cases? If not, the PGS LRs are well calibrated (or at least not systematically overstated).[369]

The point at which enough well-designed studies have been conducted to answer these questions is not easily defined. Simply counting the number of studies with the word "validity" in their title only scratches the surface. Did the studies include large numbers of contributor and noncontributor pairs? Which performance metrics were used? Is the case within the range of conditions covered in the studies? In 2016, the President's Council of Advisors on Science and Technology reported that PGS programs "clearly represent a major improvement over purely subjective interpretation,"[370] but that the studies to that point, which came from the software developers rather than from completely separate researchers,[371] "have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA)."[372] Additional studies of as many as 10-person mixtures followed.[373] Almost invariably, courts have upheld the admissibility of

369. Duncan Taylor et al., *Testing Likelihood Ratios Produced from Complex DNA Profiles*, 16 Forensic Sci. Int'l: Genetics 165 (2015), https://doi.org/10.1016/j.fsigen.2015.01.008. Simulating enough noncontributor pairs to test LRs of very large magnitudes can be challenging. For one strategy, see Duncan Taylor et al., *Importance Sampling Allows H_d True Tests of Highly Discriminating DNA Profiles*, 27 Forensic Sci. Int'l: Genetics 74 (2017), https://doi.org/10.1016/j.fsigen.2016.12.004.

370. PCAST, *supra* note 105, at 79.

371. The report evinced concern over the involvement of the developers of the software in the validity studies of their products: "Appropriate evaluation of the proposed methods should consist of studies by multiple groups, *not associated with the software developers*, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties." *Id.* (emphasis in original). The issue of developer involvement is discussed in *Gissantaner*, 990 F.3d 457 (6th Cir. 2021); *United States v. Lewis*, 442 F. Supp. 3d 1122, 1147–49 (D. Minn. 2020) (Rep. & Recommendation); *People v. Wakefield*, 195 N.E.3d 19 (N.Y. 2022); *People v. Williams*, 147 N.E.3d 1131 (N.Y. 2020); and other cases.

372. PCAST, *supra* note 105, at 80. The report added that "[t]he two most widely used methods (STRMix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem. Specifically, these methods appear to be reliable for three-person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method." *Id.*

373. *E.g.*, Michael S. Adamowicz et al., *Internal Validation of MaSTR™ Probabilistic Genotyping Software for the Interpretation of 2–5 Person Mixed DNA Profiles*, 13 Genes 1429 (2022), https://doi.org/10.3390/genes13081429; Bauer et al., *supra* note 367 (TrueAllele); Jo-Anne Bright et al., *Internal Validation of STRmix™—A Multi-Laboratory Response to PCAST*, 34 Forensic Sci. Int'l: Genetics 11 (2018), https://doi.org/10.1016/j.fsigen.2018.01.003; Sara Riman et al., *Examining Performance and Likelihood Ratios for Two Likelihood Ratio Systems Using the PROVEDIt Dataset*, 16 PLoS ONE e0256714 (2021), https://doi.org/10.1371/journal.pone.0256714 (EuroForMix and STRmix applied

PGS results in the face of objections to their scientific validity or general scientific acceptance,[374] although some cases have been remanded for more complete hearings.[375]

Presentation of LRs. Whether likelihood ratios result from categorical matching or from PGS programs, courts may need to consider the manner in which the ratios are presented at trial. The legal concerns are avoiding overstating what science has to offer (an aspect of Rules 403 and 702) and preventing unfair prejudice or confusion (Rule 403). Like random-match probabilities, LRs are easily transposed,[376] and it can be argued that large numbers might be overvalued.[377] With random-match probabilities, we saw that courts have reasoned that the possibility of transposition does not justify a blanket rule of exclusion.[378] Appellate courts have not addressed the issue for LRs,[379] but it is not obvious

to the same mixture profiles). But in 2021, a small group of scientists and other personnel working at NIST released a draft of a "scientific foundation review" that stated "[c]urrently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems." Butler et al., *supra* note 105, at 6, 75.

374. *E.g.*, *Gissantaner*, 990 F.3d 457 (STRmix); *Lewis*, 442 F. Supp. 3d 1122 (STRmix); People v. Davis, 290 Cal. Rptr. 3d 661 (Cal. Ct. App. 2022) (STRmix); State v. Simmer, 935 N.W.2d 167 (Neb. 2019) (TrueAllele); *Wakefield*, 195 N.E.3d 19 (TrueAllele); *cf.* United States v. Williams, 382 F. Supp. 3d 928 (N.D. Cal. 2019) (LR inadmissible because "Bullet" PGS program was only validated for up to four-person mixtures, and it could not be shown that the mixture in question did not come from five individuals).

375. *E.g.*, People v. Wortham, 180 N.E.3d 516 (N.Y. 2021) (New York City Office of the Chief Medical Examiner's program, FST).

376. For a few examples, see *People v. Pike*, 53 N.E.3d 147, 167 (Ill. App. Ct. 2016) ("estimates how much more likely it is that the suspect is the source of the evidence than it is that the evidence originated from a randomly selected member of the population unrelated to the suspect"); *State v. Pickett*, 246 A.3d 279 (N.J. App. 2021) ("a statistic measuring the probability that a given individual was a contributor to the sample against the probability that another, unrelated individual was the contributor"); *Commonwealth v. Treiber*, 121 A.3d 435, 445 (Pa. 2015) ("the hair was 1,000 times more likely to have come from appellant's dog than any other dog"). These statements mischaracterize the likelihood ratio as the relative probability of a source hypothesis given the DNA data rather than the relative probability of the data given the hypotheses. Kaye et al., *supra* note 193, § 14.2.2.

377. Some commentary maintains that likelihood ratios are inherently prejudicial. *E.g.,* William C. Thompson, *DNA Evidence in the O.J. Simpson Trial*, 67 U. Colo. L. Rev. 827, 855–56 (1996); *see also* Richard C. Lewontin, *Population Genetic Issues in the Forensic Use of DNA*, *in* 1 Modern Scientific Evidence: The Law and Science of Expert Testimony § 17–5.0, at 703–05 (David L. Faigman et al. eds., 1st ed. 1998).

378. *See* section titled "Frequencies, Probabilities, and Prejudice" above.

379. Trial court rulings admitting likelihood ratios over the unfair prejudice objection were upheld in *State v. Ayers*, 68 P.3d 768, 778 (Mont. 2003), and *State v. Watkins*, 648 S.W.3 235, 265 (Tenn. Crim. App. 2021).

why the rule would be different.[380] The usual expectation is that "[a] district court concerned that the jury might misunderstand what the likelihood ratio means could require advocates to describe it in a way that will not generate unfair prejudice or mislead the jury."[381]

But what would that be? It has been suggested that experts should cease characterizing LRs as statements that "a match is X times more/less probable than a coincidence,"[382] for that formulation seems to invite the misunderstanding that the LR is the odds in favor of the contributor hypothesis. Better explanations of the LR would clarify that the value pertains to the probability of the evidence (the STR data) rather than the probability of the conclusion (who the contributors are). An LR of *x* means that the DNA data are *x* times more likely if the hypothesis that includes the defendant as a contributor is correct than if the hypothesis that only other people are contributors is correct. One textbook for practitioners advises that "[a]n example of suitable phrasing (where the LR = one billion) is 'The evidence is a billion times more likely if the person of interest is a contributor than if a random, unrelated person is a contributor.'"[383]

Even with the correct conditional phrasing, however, the number may be puzzling. Factfinders are not used to thinking about ratios of probabilities. To help, experts could offer analogies. One such analogy is a partial fingerprint pattern that one in a thousand people would possess. A DNA comparison with

380. Psychological research to ascertain the effects of likelihood-ratio statements, as opposed to other quantitative and qualitative indicators of probative value, does not demonstrate that research subjects overvalue the evidence as described in questionnaires. *See* Busey, *supra* note 186; Edward K. Cheng, *The Burden of Proof and the Presentation of Forensic Results*, 130 Harv. L. Rev. F. 154, 161 (2017) ("An increasingly complex literature has emerged on lay understanding of likelihood ratios and how such quantitative information is best presented. Research thus far has yielded no easy answers . . . ."); Heidi Eldridge, *Juror Comprehension of Forensic Expert Testimony: A Literature Review and Gap Analysis*, 1 Forensic Sci. Int'l: Synergy 24 (2019), https://doi.org/10.1016/j.fsisyn.2019.03.001; Martire, *supra* note 186. For some studies in this area, see Brandon L. Garrett et al., *Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence*, 65 J. Forensic Sci. 1199 (2020), https://doi.org/10.1111/1556-4029.14323; William C. Thompson et al., *Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions*, 17 L. Probability & Risk 133 (2018), https://doi.org/10.1093/lpr.mgy012; William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 L. & Hum. Behav. 332 (2015), https://doi.org/10.1037/lhb0000134.

381. United States v. Gissantaner, 990 F.3d 457, 470 (6th Cir. 2021) (internal quotation marks and alteration omitted).

382. Thompson, *supra* note 352; *cf.* Eur. Network of Forensic Sci. Inst., Best Practice Manual for Human Forensic Biology and DNA Profiling 29 (ENFSI-DNA-BPM-03, Ver. 01. 2022) ("Caution is required when using the word 'match' in statements because it might imply 'identity'. The expert avoids any verbal statement that might imply that he/she is making an opinion on the identity of the questioned DNA.").

383. Bright & Coble, *supra* note 300, at 30.

an LR of 1,000 has the same probative value as the partial print even though the phenomena have nothing to do with each other.[384] Experts also could describe an LR as a measure of support for a scenario and add a qualitative expression such as "weak," "strong," "very strong," and so on for the degree of support.[385] Forensic statisticians originally proposed "verbal scales" for less objectively determined LRs for all kinds of forensic-science identification evidence, from fingerprints to toolmarks to handwriting, in the hope that these scales would encourage greater standardization in expressing the strength of the evidence and better comprehension of LRs.[386] The Department of Justice's Uniform Language for Testimony and Reports for PGS presents one set of verbal tags: uninformative (for LR = 1), limited support (2 to <100), moderate support (100 to <10,000), strong support (10,000 to <1,000,000), and very strong support (≥1,000,000).[387] Analysts are allowed but not required to use these tags (and no others) along with the actual value.[388] If any tag is used, then the examiner must list the full scale "to provide context for the numerical value."[389]

384. Other analogies are possible. For example, an LR of about 1,000 is what you would get if someone flipped one of two coins—either a normal, evenly balanced coin or a trick coin with heads on both sides. You do not know which coin was flipped, but the evidence is that it came up heads every time in ten tosses. The evidence—the ten heads—is about 1,000 times more probable if it was the trick coin that was tossed than if it was the normal one. The expert would have to add that the probability that the trick coin was tossed also would depend on how likely it is that the person doing the flipping would choose the trick coin instead of the normal one—a factor that is beyond the data from the experiment on the coin (tossing it ten times). Similarly, the expert would have to say that she cannot give the probability that the defendant's DNA was present. She can only report the probability *if* it was present compared to the probability *if* the alternative she considered was what had happened.

385. *E.g.*, United States v. Williams, 382 F. Supp. 3d 928, 934–35 (N.D. Cal. 2019) ("Based on SERI's [Serological Research Institute's] verbal scale, this result indicated 'very strong' support for the proposition that Elmore is a contributor to the sample.") (footnote omitted).

386. *See, e.g.,* Raymond Marquis et al., *Discussion on How to Implement a Verbal Scale in a Forensic Laboratory: Benefits, Pitfalls and Suggestions to Avoid Misunderstandings*, 56 Sci. & Just. 364 (2016), https://doi.org/10.1016/j.scijus.2016.05.009.

387. U.S. Dep't of Just., Uniform Language for Testimony and Reports for Forensic Autosomal DNA Examinations Using Probabilistic Genotyping Systems (Sept. 13, 2022) (link at https://perma.cc/8VFG-KMHB).

388. *Id.* at 3.

389. *Id.* The ULTR does not explain how dividing the continuum of possible LR values from zero to infinity into arbitrary segments provides meaningful context. As indicated in the section titled "Are Random-Match Probabilities That Are Smaller Than False-Positive Error Probabilities Irrelevant or Prejudicial?" of the third edition of this reference guide (Reference Manual on Scientific Evidence (3d ed. 2011)), some statisticians would argue that a more helpful explanation is to present an LR as the extent to which the evidence changes the odds in favor of a factual claim. *See also* Kaye & Stern, *supra* note 12, Appendix C.

DNA and RNA Typing of Bodily Fluids or Tissues

Standard DNA genotyping methods help answer the question of whose DNA is in the sample. In many cases, however, the more critical question is "how" rather than "who." For example, a defendant accused of rape might concede that there was an assault but argue that the victim's DNA found near the zipper of his pants resulted from her touching the fabric with her hands and was not from vaginal fluid, as the prosecution proposed.[390] Forensic scientists call such factual claims activity-level propositions.[391] To supply activity-level evidence, molecular biology tests for differentiating between different types of bodily fluids and tissues are beginning to be used in casework.[392] Although the sequence of base pairs in DNA throughout the body is essentially identical, different parts of the coding DNA are active in each type of specialized cell that forms a tissue. Gene "expression" occurs when part of the double-stranded DNA molecule is transcribed into messenger RNA and that mRNA is translated into a protein.[393] Some "housekeeping" proteins are in all tissues, but other proteins—and the associated mRNA transcripts—vary markedly in the extent to which they are present in different tissues. Skin cells have a different expression pattern than

390. *Cf.* Holtzclaw v. State, 448 P.3d 1134, 1148 (Okla. Ct. Crim. App. 2019) (no error for prosecutor to argue that DNA near the defendant's zipper was "transferred by the vaginal fluids" even though the DNA analyst did not make that inference); Chong Wang et al., *Body Fluid Identification by Messenger RNA Profiling in Sexual Assault*, 8 J. Forensic Sci. & Med. 118 (2022), https://doi.org/10.4103/jfsm.jfsm_54_21 ("the suspect argued that the condom was just touched by the victim").

391. R. Cook et al., *A Hierarchy of Propositions: Deciding Which Level to Address in Casework*, 38 Sci. & Just. 231 (1998), https://doi.org/10.1016/S1355-0306(98)72117-3; Bas Kokshoorn et al., *Activity Level DNA Evidence Evaluation: On Propositions Addressing the Actor or the Activity*, 278 Forensic Sci. Int'l 115 (2017), https://doi.org/10.1016/j.forsciint.2017.06.029. To address the activity-level questions, forensic-science researchers and practitioners would like to bring to bear, in a standardized and systematic way, scientific studies of transfer, persistence, prevalence, and recovery of DNA. Duncan Taylor & Bas Kokshoorn, Forensic DNA Trace Evidence Interpretation: Activity Level Propositions and Likelihood Ratios (2023); Peter Gill et al., *DNA Comm'n of the Int'l Soc'y for Forensic Genetics: Assessing the Value of Forensic Biological Evidence—Guidelines Highlighting the Importance of Propositions. Part II: Evaluation of Biological Traces Considering Activity Level Propositions*, 44 Forensic Sci. Int'l: Genetics 102186 (2020), https://doi.org/10.1016/j.fsigen.2019.102186.

392. Andrea Patrizia Salzmann et al., *mRNA Profiling of Mock Casework Samples: Results of a FoRNAP Collaborative Exercise*, 50 Forensic Sci. Int'l: Genetics 102409 (2021), https://doi.org/10.1016/j.fsigen.2020.102409; Titia Sijen & SallyAnn Harbison, *On the Identification of Body Fluids and Tissues: A Crucial Link in the Investigation and Solution of Crime*, 12 Genes 1728, § 4.1 (2021), https://doi.org/10.3390/genes12111728 ("RNA typing has been applied regularly at the Netherlands Forensic Institute since 2012"). This section discusses mRNA testing, but the related proteins themselves also can be analyzed with mass spectrometry. *See, e.g.,* Forensic Proteomics: Protein Identification and Profiling (Eric D. Merkley ed., 2019).

393. *See* section titled "Genes and Gene Products" above.

liver cells, for example.[394] Researchers have identified certain mRNA transcripts that are characteristic of forensically relevant fluids and thus can serve as markers for those fluids.

The mRNA in a biological sample can be extracted along with the DNA.[395] Using the technologies for PCR amplification and capillary electrophoresis,[396] or sequencing[397] (see section above titled "What Are DNA Polymorphisms and How Are They Detected?"), the laboratory can try to determine the fluid or tissue in the trace-evidence sample. Other RNA transcripts than mRNA also have been proposed for fluid and tissue identification, as has testing for a chemical change to the DNA molecule that silences gene expression.[398]

Twins

Identical twins originate from a single fertilized egg cell (a zygote). Instead of multiplying to develop into a single embryo, fetus, and child, the developing set of cells breaks up into two or more masses that grow into separate children.[399] Fraternal twins come from two different eggs fertilized by different sperm; those twins have many differences in their genomes. But in monozygotic twins, all the cells of the progeny come from the original zygote with a single genome. To the extent that the original chromosomes are faithfully copied in every cell

394. *See* Human Protein Atlas, The Tissue Section—Tissue-Based Map of the Human Proteome, https://perma.cc/5UW9-ULHN.

395. Amy D. Roeder & Cordula Haas, *Body Fluid Identification Using mRNA Profiling*, *in* Forensic DNA Typing Protocols 13 (William Goodwin ed., 2016), https://doi.org/10.1007/978-1-4939-3597-0.

396. Patricia Pearl Albani & Rachel Fleming, *Developmental Validation of an Enhanced mRNA-based Multiplex System for Body Fluid and Cell Type Identification*, 59 Sci. & Just. 217 (2019), https://doi.org/10.1016/j.scijus.2019.01.001; Tiffany R. Layne et al., *Rapid Microchip Electrophoretic Separation of Novel Transcriptomic Body Fluid Markers for Forensic Fluid Profiling*, 13 Micromachines 1657 (2022), https://doi.org/10.3390/mi13101657. To use a PCR-CE system, the mRNA first must be "reverse transcribed" into a DNA segment.

397. Cordula Haas et al., *Forensic Transcriptome Analysis Using Massively Parallel Sequencing*, 52 Forensic Sci. Int'l: Genetics 102486 (2021), https://doi.org/10.1016/j.fsigen.2021.102486; E. Hanson et al., *Messenger RNA Biomarker Signatures for Forensic Body Fluid Identification Revealed by Targeted RNA Sequencing*, 34 Forensic Sci. Int'l: Genetics 206 (2018), https://doi.org/10.1016/j.fsigen.2018.02.020.

398. Changes to DNA that do not alter the sequence of nucleotides are called "epigenetic." The attachment of a small chemical group (a methyl group) on a gene stops it from producing proteins. The pattern of methylation of the DNA is different in different tissues. Athina Vidaki & Manfred Kayser, *Recent Progress, Methods and Perspectives in Forensic Epigenetics*, 37 Forensic Sci. Int'l: Genetics 180 (2018), https://doi.org/10.1016/j.fsigen.2018.08.008.

399. The monozygotic twinning rate is about 1 pregnancy per 250. Kurt Benirschke, *Multiple Gestation: The Biology of Twinning*, *in* Creasy and Resnik's Maternal-Fetal Medicine: Principles and Practice 53, 53 (Robert K. Creasy et al. eds., 7th ed. 2014).

division, monozygotic twins are genetically identical.[400] Forensic STR analysis, which considers only tiny slices of the entire genome, cannot tell one from the other.

This limitation has thwarted prosecutions in which both twins were believed to have committed the crime[401] or in which an "evil twin" defense to an alleged crime committed by one twin was offered.[402] But variations in observable characteristics detectable by molecular tests related to DNA do exist.[403] This section discusses one of them that has made an initial appearance in criminal and civil litigation.[404] It has been proffered as "a realistic option, fit for practical forensic casework"[405] that can support claims that one identical twin rather than the other is the source of trace-evidence DNA.[406]

Point mutations accumulate throughout life as cells split, giving rise to more and more distinguishable lines of cells. Even before birth, many "identical"

400. *See supra* note 25 & accompanying text.

401. In perhaps the most dramatic of these cases, German authorities did not bring charges in a "massive jewelry heist" because they could not decide which twin to charge. The German twins sent a message that they were "proud of the German constitutional state and gave it their thanks." *Twins Suspected in Spectacular Jewelry Heist Set Free*, Speigel Online Int'l, Mar. 19, 2009, https://perma.cc/W6D8-GYH9, last visited Mar. 9, 2023.

402. Alison Gee, *Twin DNA Test: Why Identical Criminals May No Longer Be Safe*, BBC News Mag., Jan. 15, 2014, https://perma.cc/QM7N-DWWM; Richard Willing, *Identical Twins Complicate Use of DNA Testing*, USA Today, Nov. 30, 2004, at A03.

403. There are differences in which genes are expressing proteins ("epigenetic" differences). Mario F. Fraga et al., *Epigenetic Differences Arise During the Lifetime of Monozygotic Twins*, 102 Proc. Nat'l Acad. Sci. 10604 (2005), https://doi.org/10.1073/pnas.0500398102. Forensic-science researchers have begun to study how to use this variation on samples of the same type of tissue from two twins. *E.g.*, Benjamin Planterose Jiménez et al., *Equivalent DNA Methylation Variation Between Monozygotic Co-Twins and Unrelated Individuals Reveals Universal Epigenetic Inter-Individual Dissimilarity*, 22 Genome Biology 18 (2021), https://doi.org/10.1186/s13059-020-02223-9 ("may one day allow universal epigenetic fingerprinting, which for instance is relevant in forensics for differentiating MZ twin individuals"); Athena Vidaki et al., *Epigenetic Discrimination of Identical Twins from Blood Under the Forensic Scenario*, 31 Forensic Sci. Int'l: Genetics 67 (2017), https://doi.org/10.1016/j.fsigen.2017.07.014. In addition, there are differences in the number of copies of a particular gene present in the genomes of different twins ("copy number variants"). Carl E.G. Bruder et al., *Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles*, 82 Am. J. Hum. Genetics 763 (2008), https://doi.org/10.1016/j.ajhg.2007.12.011. Furthermore, differences in mitochondrial sequences have been reported. Lijuan Yuan et al., *Identification of the Perpetrator Among Identical Twins Using Next Generation Sequencing Technology: A Case Report*, 44 Forensic Sci. Int'l: Genetics 102167 (2020), https://doi.org/10.1016/j.fsigen.2019.102167.

404. Burkhart Rolf & Michael Krawczak, *The Germlines of Male Monozygotic (MZ) Twins: Very Similar, But Not Identical*, 50 Forensic Sci. Int'l: Genetics 102408 (2021), https://doi.org/10.1016/j.fsigen.2020.102408.

405. Michael Krawczak et al., *Distinguishing Genetically Between the Germlines of Male Monozygotic Twins*, 14 PLoS Genetics e1007756 (2018), https://doi.org/10.1371/journal.pgen.1007756.

406. The technique has been used several times in Europe. *Id.*; Netherlands Forensic Institute, NFI Reports on Distinguishing Twins Using DNA Analysis, Oct. 19, 2022, https://perma.cc/XWX8-83Y8.

twins differ at some single base pairs (out of billions).[407] After the embryo splits into twins, cells in each twin migrate and differentiate into distinct tissue types. If the mutation event occurs after twinning, it will be present only in one twin.[408] If it occurs early, before the cells have specialized into all the cell types (nerve cells, blood cells, skin cells, germline cells, and so on), it may be perpetuated in multiple cell types and tissues. And if it occurs farther down the road, it will appear only in the tissue type of the first mutated cell.

To find distinguishing point mutations, a laboratory can perform whole-genome sequencing with an MPS method[409] on a sample of blood from each twin. The twins are likely to have a small number of discordant base pairs.[410] Knowing the few discrepancies between the twins, the laboratory goes back to the trace-evidence DNA. Suppose the trace-evidence DNA came from a bloodstain. Sanger sequencing of that DNA at the distinguishing SNPs will show which twin's SNPs are also present in the trace-evidence DNA.

In this example, blood DNA has been compared to blood DNA. Suppose instead that the trace-evidence DNA is from semen, not blood, and the comparison DNA is from swabs of insides of the cheeks (the buccal mucosa) of the twins. The two types of collected cells developed from different parts of the embryo, and this fact complicates the analysis. Two or three weeks after fertilization, a set of cells in the embryo are slated to become germ cells (sperm or ova) in an event known as primordial germ cell specification (PGCS). Whether a mutation is propagated in sperm, in the mucosa, or in both—and whether these outcomes are present in one twin, the other twin, or both—depends on when (relative to PGCS and twinning) and where (within the embryo) they occurred.[411] With cross-tissue comparisons, the technique thus requires establishing that at least some of the different SNPs in the twins' tissues are "twin-specific on the

407. Hakon Jonsson et al., *Differences Between Germline Genomes of Monozygotic Twins*, 53 Nature Genetics 27, 27 (2021), https://doi.org/10.1038/s41588-020-00755-1 ("monozygotic twins differ on average by 5.2 early developmental mutations and . . . approximately 15% of monozygotic twins have a substantial number of these early developmental mutations specific to one of them"); Rui Li et al., *Somatic Point Mutations Occurring Early in Development: A Monozygotic Twin Study*, 51 J. Med. Genetics 28 (2014), https://doi.org/10.1136/jmedgenet-2013-101712 (based on different SNPs found in 66 monozygotic twin pairs, "it is likely that each individual carries approximately over 300 postzygotic mutations in the nuclear genome of white blood cells that occurred in the early development").

408. Conceivably, the same mutation could occur independently after twinning at the very same point within the billions of base pairs in a cell in the other twin. In that case, the single nucleotide change would not distinguish between the twins, but such a duplicated mutation is extremely unlikely.

409. *See supra* section titled "Sequencing and SNPs."

410. To prune away sequencing errors with MPS, the laboratory can apply a slower but more accurate method such as Sanger sequencing to the small number of candidate variants.

411. *See* Jonsson et al., *supra* note 407.

one hand, but not (too) tissue-specific on the other."[412] In the semen-versus-cheek-swab situation, the premise is that the mutations occurred in early cells that gave rise to both progenitor germ cells and epithelial cells.

The evidentiary issues associated with this method of breaking the STR tie between twins include the validity of the sequencing methods;[413] the adequacy of the statistical model for calculating a pertinent likelihood ratio;[414] and the laboratory's correct execution of the sequencing and statistical procedures. Also, a criminal-procedure issue would arise if one or both of the twins refused to give DNA samples. Compelling suspects to give bodily samples (breath, blood, saliva, hair, and fingerprint scrapings) is not unusual,[415] and warrants, nontestimonial orders, or subpoenas for DNA samples from third parties have been issued.[416] But these cases contemplated forensic STR or VNTR profiling, which reveals limited medical or other socially significant information.[417] Whole genome sequencing (WGS) reveals copious genetic information. Even if the only part of the genome that the government cares about and uses is a handful of SNPs that are useful only for differentiating between the two twins, it could be argued that compelling the twins to submit to the extraction of their blood or saliva for WGS is an unreasonable search or seizure.[418]

412. Rolf & Krawczak, *supra* note 404. This wrinkle resulted in the exclusion of whole genome sequencing evidence in *Commonwealth v. McNair*, No. 8414CR10768 (Mass. Suffolk Cnty. Super. Ct., Apr. 11, 2017), *pet. for interlocutory rev. denied*, No. SJ-2017-0186 (Supreme Jud. Ct., Suffolk Cnty., July 27, 2017). The memorandum opinion is reprinted in David H. Kaye, *The Gordian Knot in* Commonwealth v. McNair: *What Did the Court Decide About* Daubert*?*, Forensic Sci., Stat. & L., Mar. 11, 2019, https://perma.cc/3LJJ-P4K4.

413. *See supra* section titled "Sequencing methods."

414. Michael Krawczak et al., *Distinguishing Genetically Between the Germlines of Male Monozygotic Twins*, 14 PLoS Genetics e1007756 (2018), https://doi.org/10.1371/journal.pgen.1007756; David H. Kaye, *Identical Twins, Different Tissues, Disparate DNA*, Forensic Sci., Stat. & L., Mar. 7, 2019, https://perma.cc/49AP-Q4WL.

415. Doe v. Senechal, 725 N.E.2d 225 (Mass. 2000) (court-ordered buccal swab to test whether a member of the staff of a residential treatment facility for mentally ill adolescents fathered the child of a patient is a reasonable search and seizure); *In re* Non-testimonial Identification Order Directed to R.H., 762 A.2d 1239 (Vt. 2000) (saliva sampling by court order based on reasonable suspicion).

416. Bill v. Brewer, 799 F.3d 1295 (9th Cir. 2015) (warrant for police officers' DNA to exclude them as sources of crime-scene DNA); *In re* G.B., 139 A.3d 885 (D.C. 2016) (warrant for a buccal swab from the victim of a stabbing for testing of DNA from blood droplets and stains at the apartment where the victim allegedly was stabbed). *But cf.* Commonwealth v. Kostka, 31 N.E.3d 1116 (Mass. 2015) (order to twin to provide a saliva sample to determine whether he was a fraternal as opposed to an identical twin was unreasonable where there was no showing of how definitive the profiling of a mixed sample from the victim's fingernail was and no showing that defendant would raise an evil twin defense; and where fingerprint evidence could have excluded even an identical twin).

417. *See supra* note 39.

418. *Cf.* section titled "Investigative Genetic Genealogy" below. To forestall arguments over the validity or weight of cross-tissue comparisons, the prosecution in a rape case would need to

Investigative DNA Analysis

Much of this guide has presented DNA analysis as courtroom evidence that a criminal defendant is associated with a crime. The remaining sections describe DNA analysis from the perspective of police who have yet to zero in on a suspect. How can forensic scientists use DNA from a crime scene or a victim as a clue that could lead police to some actual suspects? Computerized databases of DNA profiles serve this function by allowing investigators to compare an unknown profile to millions of profiles from known individuals for "cold hits." These law-enforcement intelligence databases are the subject of the next section. Police also can take advantage of large databases of other information on individual genomes maintained by companies to help members of the public locate possible relatives. The section below titled "Investigative Genetic Genealogy" describes the use of this genomic data to locate suspects indirectly, by first identifying possible relatives of the unknown source of the crime-scene DNA. A final section considers trace-evidence DNA, not purely as a token of individual identity or a step in constructing a family tree, but rather as an indication of biogeographic origin or external traits.

Offender and Suspect Database Searches

Law-enforcement databases and databanks

In 1989, Virginia became the first state to collect DNA from individuals convicted of certain crimes for a statewide DNA database—a set of records of DNA profiles that could be searched to see whether any of them matched a profile from biological evidence associated with a crime. Other states established similar systems,[419] and in 1998, the FBI launched software and hardware—the Combined DNA Index System (CODIS)—to allow nationwide searches to generate

secure semen samples from the twins. The balance of interests that would determine the constitutional reasonableness of compelling a twin to give such a sample would tilt further in the reluctant twin's favor.

419. Although state and federal statutes authorizing and governing DNA databases for law enforcement are now pervasive, it is not obvious that the laws preclude a local police agency from maintaining its own records of DNA profiles from suspects they encounter or from biological evidence that they collect. "Rapid DNA" instruments and companies marketing software for database management have encouraged this practice. *See supra* section titled "Miniaturization for rapid capillary electrophoresis." The "rogue databases" operating outside the constraints and quality assurance requirements of statutorily established systems have sparked concern. Murphy, *supra* note 287, at 180–88; N.Y. City Bar Ass'n Crim. Courts Comm., Crim. Just. Operations Comm., and Mass Incarceration Task Force, Curbing Unregulated Local DNA Indexing (Apr. 16, 2021).

an extremely valuable investigative lead: the name of an individual whose DNA, very likely, was left at the crime scene.[420]

CODIS has three layers. At the bottom, local governmental laboratories store the DNA profiles they obtain in a Local DNA Index System (LDIS). In the middle, a State DNA Index System (SDIS) holds these profiles and additional ones generated by statewide laboratories. State law specifies which profiles can be included. A National DNA Index System (NDIS) sits at the top. It houses SDIS profiles that meet federal requirements plus DNA profiles from federal laboratories. Approved law-enforcement laboratories can conduct database searches at any level—local, state, or national.[421]

These CODIS databases have four main sets of records: a convicted offender index; an arrestee index; a detainee index;[422] and a forensic index (of profiles from DNA from crime scenes or left on victims).[423] Matches can occur between either the convicted offender, arrestee, or detainee indexes and the forensic index. Cold hits also can occur within the forensic index, linking crimes to each other. The underlying samples (or just the DNA extracts) also are retained, making the system a databank as well as an electronic database.[424] Finally, the system as a whole includes "population databases." These are anonymized research databases used to estimate allele frequencies that go into the equations

420. A pilot project had been started in 1990. U.S. Dep't of Just., Office of the Inspector General, Combined DNA Index System Operational and Laboratory Vulnerabilities, Audit Report 06–32 (2006). The DNA Identification Act of 1994, 34 U.S.C. §§ 40702–40703 (later amended in various respects), authorized the FBI to create a national database. At the outset of the fully operational system, only nine states participated, and the federal government did not collect DNA from federal offenders. Nathan James, Cong. Reference Serv., DNA Testing in Criminal Justice: Background, Current Law, Grants, and Issues 3 (2014) (CRS Rep. R41800).

421. On the requirements for uploading a crime-scene profile to NDIS so that it will be searched against the offender and arrestee databases, see *Cowels v. FBI*, 936 F.3d 62 (1st Cir. 2019) (FBI's refusal to accept a sample that a state court ordered uploaded to NDIS at the request of defendants, who had been convicted but were granted a new trial, was not arbitrary and capricious under the Administrative Procedure Act where it was not clear that the sample was connected to the crime). The scope of the information maintained at each level of the hierarchy of databases varies. LDIS has names and crime information that are not uploaded to SDIS and NDIS.

422. Detainees are "non-United States persons who are detained under the authority of the United States." 28 C.F.R. § 28.12(b) (2023). "Non-United States persons" means "persons who are not United States citizens and who are not lawfully admitted for permanent residence." *Id*. DNA is not routinely collected from "(1) [a]liens lawfully in, or being processed for lawful admission to, the United States; (2) [a]liens held at a port of entry during consideration of admissibility and not subject to further detention or proceedings; or (3) [a]liens held in connection with maritime interdiction." *Id*. On the rationale for these exceptions, see 85 Fed. Reg. 13483 (Mar. 9, 2020).

423. It also has profiles from unidentified remains and from samples voluntarily provided by relatives of missing persons. Butler, *supra* note 43, at 236.

424. *Id*. at 214–15; United States v. Kriesel, 720 F.3d 1137 (9th Cir. 2013) (describing the reasons for sample retention and holding that Fed. R. Crim. 41(g) does not require the government to return the blood sample after an offender has completed his sentence and term of supervised release).

for genotype frequencies, random-match probabilities, or likelihood ratios.[425] They come from other sources and are not searched for cold hits.[426]

Law-enforcement DNA databases and databanks have provoked considerable litigation. Fourth Amendment challenges to compulsory DNA collection from convicted offenders almost always failed,[427] although opinions were divided as to the appropriate doctrinal route around normal warrant and probable cause requirements,[428] and several vigorous dissenting opinions emerged.[429] Expanding the scope of the databases to include profiles from individuals who were merely indicted or arrested produced a split of authority in the lower courts. Eventually, the Supreme Court, in *Maryland v. King*,[430] narrowly upheld a state statute requiring all custodial arrestees to submit to DNA sampling as part of the booking process for certain violent crimes and burglary.[431]

425. These quantities are discussed *supra* in the sections titled "Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing" and "(1) Definition of the likelihood ratio".

426. Butler, *supra* note 43, at 215. The failure to recognize the difference between searchable records for possible cold hits and de-identified databases for population frequency research led the state of Texas to destroy its public health system's repository of newborn blood samples. David H. Kaye, *Up in Smoke: 5 Million Neonatal Blood Samples Incinerated*, Forensic Sci., Stat. & L., July 27, 2014, https://perma.cc/2LN2-NQXL. On law-enforcement access to newborn samples collected for genetic-disease screening, see Natalie Ram, *America's Hidden National DNA Database*, 100 Tex. L. Rev. 1253 (2022).

427. *See* United States v. Kincade, 379 F.3d 813 (9th Cir. 2004) (en banc); Robin Cheryl Miller, *Validity, Construction, and Operation of State DNA Database Statutes*, 76 A.L.R. 5th 239 (2000).

428. *See* Banks v. United States, 490 F.3d 1178, 1183 (10th Cir. 2007) ("our sister circuits have taken different analytical routes to analyzing DNA-indexing statutes" and "our own precedents are divided"); Padgett v. Donald, 401 F.3d 1273, 1278 (11th Cir. 2005) ("Each circuit to address the question has upheld the constitutionality of DNA profiling statutes, but the circuits have disagreed on whether to do so through the special needs analysis or through the traditional balancing test.").

429. *Kincade*, 379 F.3d at 842 & 871 (dissenting opinions).

430. 569 U.S. 435 (2013).

431. *King* has been extended to uphold other database laws that encompass more crimes, that do not automatically remove profiles and samples if no conviction follows the arrest, and that can be used for "familial searching." Haskell v. Harris, 745 F.3d 1269 (9th Cir. 2014) (en banc) (per curiam); Haskell v. Brown, 317 F. Supp. 3d 1095 (N.D. Cal. 2018); People v. Buza, 413 P.3d 1132 (Cal. 2018) (but defendant's alleged crime fell within the "serious" category of *King*); *cf.* State v. Medina, 102 A.3d 661 (Vt. 2014) (striking down post-arraignment, pre-conviction, DNA sampling under the Vermont constitution). For academic debate on the reach of *King* and the Fourth Amendment logic of the majority and the dissent, compare Erin Murphy, *License, Registration, Cheek Swab: DNA Testing and the Divided Court*, 127 Harv. L. Rev. 161 (2013), with David H. Kaye, Maryland v. King: *Per Se Unreasonableness, the Golden Rule, and the Future of DNA Databases*, 127 Harv. L. Rev. Forum 39 (2013). *See also* David H. Kaye, *Why So Contrived? DNA Databases After* Maryland v. King, 104 J. Crim. L. & Criminology 535 (2014); Tracey Maclin, Maryland v. King: Terry v. Ohio *Redux*, 2013 Sup. Ct. Rev. 359; Andrea Roth, Maryland v. King *and the Wonderful, Horrible DNA Revolution in Law Enforcement*, 11 Ohio St. J. Crim. L. 295 (2013).

Probative value of database hits

If the DNA profile from a trace-evidence sample matches a profile in an arrestee, detainee, or offender database, the person identified by the cold hit will become the target of the investigation. Prosecution may follow. These database-trawl cases can be contrasted with traditional "confirmation cases" in which the defendant already was a suspect and the DNA testing provided additional evidence. In confirmation cases, statistics such as the estimated frequency of the matching DNA profile in various populations, the equivalent random-match probabilities, or likelihood ratios can be used to indicate the probative value of the DNA match.[432]

In trawl cases, however, an additional question arises: Does the fact that the defendant was selected for prosecution by trawling require some adjustment to the usual statistics? The legal issues are twofold. First, is a particular quantity—be it the unadjusted random-match probability or some adjusted probability—scientifically valid (or generally accepted) in the case of a database search? If not, it must be excluded under the *Daubert* (or *Frye*) standards. Second, is the statistic irrelevant or unduly misleading? If so, it must be excluded under the rules that require all evidence to be relevant and not unfairly prejudicial. To clarify, we summarize the statistical literature on this point. Then, we describe the case law.

The statistical analyses of adjustment. All statisticians agree that, in principle, the search strategy affects the probative value of a DNA match. One group describes and emphasizes the impact of the database match on the hypothesis that the database does not contain the source of the crime-scene DNA. This is a "frequentist" view. It asks how frequently searches of innocent databases—those for which the true source is someone outside the database—will generate cold hits. From this perspective, trawling is a form of "data mining" that produces a "selection effect" or "ascertainment bias."[433] If we pick a lottery ticket at random, the probability p that we have the winning ticket is negligible. But if we search through all the tickets, sooner or later we will find the winning one. And even if we search through some smaller number N of tickets, the probability of picking a winning ticket is no longer p, but Np.[434]

432. On the computation and admissibility of such statistics, see *supra* sections titled "Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing" and "Likelihood ratio (LR)" (in the section titled "Mixtures").

433. *See* Kaye & Stern, *supra* note 12.

434. The analysis of the DNA database search is more complicated than the lottery example suggests. In the simple lottery, there was exactly one winner. The trawl case is closer to a lottery in which we hold a ticket with a winning number, but it might be counterfeit, and we are not sure how many counterfeit copies of the winning ticket were in circulation when we bought our N tickets.

Likewise, if DNA from *N* innocent people is examined to determine if any of them match the crime-scene DNA, then the probability of a match in this group is not *p*, but some quantity that could be as large as *Np*. This type of reasoning led the 1996 NRC committee to recommend that "[w]hen the suspect is found by a search of DNA databases, the random-match probability should be multiplied by *N*, the number of persons in the database."[435] The 1992 committee[436] and the FBI's former DNA Advisory Board[437] took a similar position.

No one questions the mathematics that show that when the database size *N* is very small compared with the size of the population, *Np* is an upper bound on the expected frequency with which searches of databases will incriminate innocent individuals when the true source of the crime-scene DNA is not represented in the databases. The Bayesian school of thought, however, suggests that the frequency with which innocent databases will be falsely accused of harboring the source of the crime-scene DNA is basically irrelevant. The question of interest to the legal system is whether the one individual whose database DNA matches the trace-evidence DNA is the source of that trace. As the size of a database approaches that of the entire population, finding one and only one matching individual should be more, not less, convincing evidence against that person. Thus, instead of looking at how surprising it would be to find a match in a large group of innocent suspects, this school of thought asks how much the result of the database search enhances the probability that the individual so identified is the source. The database search is actually more probative than the confirmation search because the DNA evidence in the trawl case is much more extensive. Trawling through large databases excludes millions of people, thereby reducing the number of people who might have left the trace evidence if the suspect did not. This additional information increases the likelihood that the defendant is the source, although the effect is indirect and generally small.[438]

435. NRC II, *supra* note 7, at 161 (Recommendation 5.1).

436. The first NRC committee wrote that "[t]he distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important." It used the same *Np* formula in a numerical example to show that "[t]he chance of finding a match in the second case is considerably higher, because one . . . fishes through the databank, trying out many hypotheses." NRC I, *supra* note 6, at 124. Rather than proposing a statistical adjustment to the match probability, however, that committee recommended using only a few loci in the databank search, then confirming the match with additional loci, and presenting only "the statistical frequency associated with the additional loci." *Id*. at 124 tbl. 1.1.

437. DNA Advisory Bd., Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated from Pertinent Population Database(s) (July 2000), https://perma.cc/JQ4C-6ZGR.

438. When the size of the database approaches the size of the entire population, the effect is large. Also, finding that only one individual in a large database has a particular profile raises the probability that this profile is very rare, further enhancing the probative value of the DNA evidence.

Of course, when the cold hit is the only evidence against the defendant, the total package of evidence in the trawl case is less than in the confirmation case. Nonetheless, the Bayesian treatment shows that the DNA part of the total evidence is more powerful in a cold-hit case because this part of the evidence is more complete than when the search for matching DNA is limited to a single suspect. This reasoning suggests that the random-match probability (or, equivalently, the frequency p in the population) understates the probative value of the unique DNA match in the trawl case. And if this is so, then Np is misleadingly large, and the unadjusted random-match probability or frequency p can be used as a conservative indication of the probative value of the finding that, of the many people in the database, only the defendant matches.[439]

Judicial opinions on adjustment. The need for an adjustment has been vigorously debated in the statistical literature, and to a lesser extent in the legal literature.[440] The dominant view in the journal articles is that the random-match probability, frequency, or likelihood ratio need not be inflated to protect the defendant—at least, not as a matter of mathematics.[441] The major opinions to

439. This analysis was developed by David Balding and Peter Donnelly. For informal expositions, see, for example, Peter Donnelly & Richard D. Friedman, *DNA Database Searches and the Legal Consumption of Scientific Evidence*, 97 Mich. L. Rev. 931 (1999); David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N. Car. L. Rev. 425 (2009); Simon Walsh et al., *DNA Intelligence Databases*, *in* Forensic DNA Evidence Interpretation 429, 443–44 (John Buckleton et al. eds., 2d ed. 2016); Simon Walsh & John Buckleton, *DNA Intelligence Databases*, *in* Forensic DNA Evidence Interpretation 439, 465–68 (John Buckleton et al. eds., 2005). For a related analysis directed at the average probability that an individual identified through a database trawl is the source of a crime-scene DNA sample, see Yun S. Song et al., *Average Probability That a "Cold Hit" in a DNA Database Search Results in an Erroneous Attribution*, 54 J. Forensic Sci. 22, 23–24 (2009), https://doi.org/10.1111/j.1556-4029.2008.00917.x. Yet another variation is in Ian Ayres & Barry Nalebuff, *The Rule of Probabilities: A Practical Approach for Applying Bayes' Rule to the Analysis of DNA Evidence*, 67 Stan. L. Rev. 1447 (2015); *cf.* John T. Wixted et al., *Calculating the Posterior Odds from a Single-match DNA Database Search*, 18 L. Probability & Risk 1 (2018), https://doi.org/10.1093/lpr/mgz001 (with commentary).

440. For citations to this literature, see Kaye, *supra* note 439; Ronald W.J. Meester & Klaas Slooten, *DNA Database Matches: A p Versus np Problem*, 46 Forensic Sci. Int'l: Genetics 102229 (2020); https://doi.org/10.1016/j.fsigen.2019.102229; Walsh et al., *supra* note 439, at 443.

441. Marjan J. Sjerps, *Probabilistic Considerations when Interpreting Database Searches and Selection Effects*, *in* Handbook of Forensic Statistics 325, 331 (David Banks et al. eds., 2021) ("[N]umerous papers have been published after Donnelly and Friedman [*supra* note 439] and [A. Philip Dawid, *Comment on Stockmarr's "Likelihood Ratios for Evaluating DNA Evidence When the Suspect Is Found Through a Database Search*," 57 Biometrics 976 (2001)], exploring the controversy with, e.g., Bayesian networks, or rewriting formulas in a more convenient way, reaching for the most part the same conclusion. . . . Thus, from a mathematical point of view, the database search controversy has been solved."). On the manner and situations in which p, Np, or some combination or variation of the two should be presented, there is less agreement. *See* ENFSI DNA Working Group, DNA Database

confront this issue agree that *p* is admissible against the defendant in a trawl case.[442] They reason that all the statistics are admissible under *Frye* and *Daubert* because there is no controversy over how they are computed. They then assume that both *p* and *Np* are logically relevant and not prejudicial in a trawl case.

Kinship trawling ("familial searching")

Normally, police trawl a DNA database to see if any recorded STR profiles match a crime-scene profile. On occasion, these trawls could lead to charges against a defendant whose DNA is not even in the databank. This can occur for identical twins, one of whom is a convicted offender whose DNA type is on file and the other who has never been typed. If the convicted twin was in prison when the crime under investigation was committed, and if the police realize that he had an identical twin, suspicion should fall on the twin. Presumably, the police would seek a sample from this twin, and at trial it would not be necessary for the prosecution to explain the roundabout process through which the defendant was identified.

In this hypothetical example, the defendant fortuitously was found because a relative's DNA led the police to him. More generally, the fact that close relatives share more alleles than other members of the same subpopulation can be exploited as a deliberate investigative tool. Rather than search for a match at all the loci in an STR profile, police could search for a near miss—a partial match that is much more probable when the partly matching profile in the database comes from a close relative than when it comes from an unrelated person.[443]

Management Review and Recommendations (Apr. 2019), accessible via European Network of Forensic Science Institutes (ENFSI) Best Practice Manuals and Forensic Guidelines, https://perma.cc/XN6B-8DAY; David H. Kaye, People v. Nelson: *A Tale of Two Statistics*, 7 Law, Probability & Risk 249 (2008), https://doi.org/10.1093/lpr/mgn005; Kaye, *supra* note 439; Meester & Slooten, *supra* note 440; Murphy, *supra* note 287, at 118; Sjerps, *supra*; Wixted et al. (and commentary), *supra* note 439.

442. People v. Nelson, 185 P.3d 49 (Cal. 2008); United States v. Jenkins, 887 A.2d 1013 (D.C. 2005). The cases are analyzed in Kaye, *supra* note 439 & note 441. More recent cases are similar. *E.g.*, People v. Turner, 476 P.3d 676 (Cal. 2020); Commonwealth v. Bizanowicz, 945 N.E.2d 356, 362–63 (Mass. 2011). In *Turner*, the California Supreme Court took a step toward the pure random-match probability approach. 476 P.3d at 693 & 693 n.12 (intimating that the *Np* statistic would not be "germane" when *p* is extremely small, but it could be admitted "in some circumstances" when "a profile's random match probability [is] relatively high"). Why or how to draw this line is not clear.

443. Frederick R. Bieber et al., *Finding Criminals Through DNA of Their Relatives*, 312 Science 1315 (2006), https://doi.org/10.1126/science.1122655. The FBI distinguishes between kinship trawling to locate an individual outside the database and "partial match" searching to locate an individual inside the database when the crime-scene profile is incomplete. FBI, *supra* note 9 (questions 25 & 27–32). The latter can incidentally produce a lead to someone outside the database, but it is not an efficient procedure for that purpose.

Analysis of Y-STRs or mtDNA from the DNA sample in the databank then could determine whether the "database inhabitant" probably is in the same paternal or maternal lineage as the unknown individual who left DNA at the scene of the crime. Apparent relatives who emerge from this process become candidates for further investigation based on nongenetic information such as age and location. (Alternatively, a Y-STR or mtDNA database could be queried at the outset for possible matches to the trace-evidence DNA. Several countries have used this strategy.[444])

Such "familial searching"[445] raises technical and policy questions. The technical and practical challenge is to devise a search strategy that detects the relatives who happen to be in the database while keeping the number of false leads to a tolerable level. For autosomal STR profiling supplemented by Y-STR or mtDNA filtering, software that computes a "kinship index" for STR profiles skims the database for first-degree relatives (parent–child and sibling relationships) based on the pattern of allele sharing and the frequency of the shared alleles.[446]

The policy questions are whether exposing relatives to the possibility of being investigated on the basis of genetic leads from their kin is fair and whether the benefits of kinship trawling outweigh the dangers of intrusions from false

444. Jianye Ge & Bruce Budowle, *Forensic Investigation Approaches of Searching Relatives in DNA Databases*, 66 J. Forensic Sci. 430 (2021), https://doi.org/10.1111/1556-4029.14615 (explaining the advantages and noting the costs of this form of kinship trawling).

445. On this term and possibly clearer alternatives to it, see David H. Kaye, *The Genealogy Detectives: A Constitutional Analysis of "Familial Searching,"* 50 Am. Crim. L. Rev. 109, 120–21 (2013). "Familial searching" became prominent in the United States with the conviction of the serial killer dubbed the Grim Sleeper, who seemed to take long breaks between murders of young women in South Los Angeles from the mid-1980s until at least 2007. Mike McPhate, *Los Angeles Man Is Convicted in 'Grim Sleeper' Serial Killer Case*, N.Y. Times, May 5, 2016. California authorities located him via the DNA profile of his son, whose profile was in the state database as the result of a 2008 arrest for firearm and drug offenses. Marisa Gerber, *The Controversial DNA Search that Helped Nab the 'Grim Sleeper' Is Winning over Skeptics*, L.A. Times, Oct. 25, 2015. A subsequent and even more highly publicized development—involving trawls of nongovernmental databases for large blocks of SNPs that reveal much more distant relationships—is discussed *infra* in the section titled "Investigative Genetic Genealogy." For clarity, the term "familial searching" should be reserved for the trawls, like the one in the Grim Sleeper case, for close relatives in law-enforcement databases.

446. The database profiles with the largest kinship-index values rise to the top. A siblingship index, a paternity index, or any other kinship index is a kind of likelihood ratio (a term explained *supra* section titled "(1) Definition of the likelihood ratio"). It contrasts the probability of observing the STR types when the profile in the database and the profile of the unknown source of the crime-scene sample have a certain relationship (for example, siblingship) to the probability of observing these types when they are unrelated. Jianye Ge et al., *Comparisons of Familial DNA Database Searching Strategies*, 56 J. Forensic Sci. 1448, 1448 (2011), https://doi.org/10.1111/j.1556-4029.2011.01867.x; Steven P. Myers et al., *Searching for First-Degree Familial Relationships in California's Offender DNA Database: Validation of a Likelihood Ratio-Based Approach*, 5 Forensic Sci. Int'l: Genetics 493 (2011), https://doi.org/10.1016/j.fsigen.2010.10.010.

leads, disruption of family harmony, and exposure of family or individual secrets, such as misattributed paternity.[447] The extent to which such concerns could support constitutional challenges to kinship trawling has been debated.[448] A few states have banned kinship trawling of their law-enforcement databases, and several jurisdictions subject it to "an approval process, specified and limited offense types, and an affirmative assertion by specific stakeholders that all leads have been exhausted."[449]

All-pairs matching within a database to verify estimated random-match probabilities

A third and final use of police intelligence databases has evidentiary implications. The section above titled "Frequencies, Probabilities, and Prejudice" explained the use of population-genetics models to estimate DNA-genotype frequencies. Large databases can be used to check these theoretical computations. Using the New Zealand national database, which consisted of six-locus profiles, for example, researchers compared every profile with every other profile.[450] At the time, there were 10,907 profiles in the database. Although this number is not huge, 10,907 profiles can be arranged to form about 59 million distinct pairs.[451] Because the theoretical random-match probability was about 1 in 50 million, if all the individuals represented in the database were unrelated, one would expect that an exhaustive comparison of the profiles for these

447. *See, e.g.,* Henry T. Greely et al., *Family Ties: The Use of DNA Offender Databases to Catch Offenders' Kin*, 34 J.L. Med. & Ethics 248 (2006), https://doi.org/10.1111/j.1748-720X.2006.00031.x; Erica Haimes, *Social and Ethical Issues in the Use of Familiar Searching in Forensic Investigations: Insights from Family and Kinship Studies*, 34 J.L. Med. & Ethics 263 (2006), https://doi.org/10.1111/j.1748-720X.2006.00032.x; Erin Murphy, *Relative Doubt: Familial Searches of DNA Databases*, 109 Mich. L. Rev. 291 (2010); Murphy, *supra* note 287, at 189–214; Sonia M. Suter, *All in the Family: Privacy and DNA Familial Searching*, 23 Harv. J.L. & Tech. 309 (2010).

448. *Compare* Murphy, *supra* note 447, and Murphy, *supra* note 287, at 208–11, *with* Kaye, *supra* 445.

449. Ge & Budowle, *supra* note 444, at 432; *see also* Tepring Piquado et al., Forensic Familial and Moderate Stringency DNA Searches: Policies and Practices in the United States, England, and Wales (2019), https://perma.cc/8HPE-V32Y.

450. Walsh & Buckleton, *supra* note 439, at 463.

451. For the people represented in the New Zealand database, about 10,907 × 10,907 pairs—such as (1,1), (1,2), (1,3), . . . , (1,10907), (2,1), (2,2), (2,3), . . . , (2,10907), . . . , (10907,1), (10907,2), (10907,3), (10907,10907)—can be formed. This amounts to almost 119 million possible pairs. But there is no point in checking the pairs (1,1), (2,2), . . . (10,907,10,907)—a profile obviously matches itself. Thus, the number of ordered pairs with different individuals is 119 million minus 10,907. Finally, ordered pairs such as (1,5) and (5,1) involve the same two people. Therefore, the number of *distinct* pairs of people is about half of 119 million—the 59 million figure in the text. More generally, an all-pairs search in a large database of size N involves $N(N–1)/2$, or about $N^2/2$ comparisons.

59 million pairs would produce only about one match. In fact, the 59 million comparisons revealed 10 matches. What could have caused this excess number of matches? Further investigation showed that eight of the pairs were twins or brothers, the ninth was a duplicate (because one person gave a sample as himself and then again pretending to be someone else), and the tenth was apparently a match between two unrelated people. This exercise thus confirmed the theoretical computation of the random-match probability. On average, the theoretical match probability was about 1/50,000,000, and the rate of matches in the unrelated pairs within the database was 1/59,000,000.

In the United States, defendants have argued that the unintuitively high number of partial matches from internal all-pairs trawls of offender databases demonstrates that the usual estimates of random-match probabilities are incorrect,[452] and they have sought discovery of the criminal-offender databases to determine whether the number of matching and partially matching pairs exceeds the predictions made with the population-genetics model.[453] An early report about partial matches (at any nine loci out of the 13 CODIS core loci) in the state database in Arizona was said to show extraordinarily large numbers of partial matches (without accounting for the combinatorial explosion in the number of comparisons in an all-pairs database search).[454] However, some

452. *E.g.*, People v. Luna, 989 N.E.2d 655, 664–65 (Ill. App. Ct. 2013) (defense expert's testimony on results of an all-pairs trawl of the Illinois state dataset).

453. United States v. Davis, 602 F. Supp. 2d 658, 681 (D. Md. 2009) (reporting results from an all-pairs search of the Maryland database ordered by a state court); Derr v. State, 73 A.3d 254 (Md. 2013) (defendant who matched at 13 loci was not entitled to have the FBI determine how many pairwise matches there would be at 9 through 13 loci; neither was he entitled to all the profiles in the national database so that his experts could conduct these pairwise comparisons); State v. Dwyer, 985 A.2d 469 (Me. 2009) (defendant was not entitled to require the state to perform an all-pairs search of its database for matches at nine or more loci); Young v. United States, 63 A.3d 1033 (D.C. 2013) (defendant who matched at 13 loci was not entitled to all-pairs searching of the national database); Jason Felch & Maura Dolan, *How Reliable Is DNA in Identifying Suspects?*, L.A. Times, July 20, 2008.

454. Felch & Dolan, *supra* note 453; David H. Kaye, *Trawling DNA Databases for Partial Matches: What Is the FBI Afraid Of?*, 19 Cornell J.L. & Pub. Pol'y 145 (2009); Murphy, *supra* note 287, at 107–08. As illustrated, *supra* note 451, an all-pairs search in a large database of size N involves nearly $N^2/2$ comparisons. For example, a database of 6 million samples gives rise to some 18,000,000,000,000 comparisons. Even with no population structure, relatives, and duplicates, and with random-match probabilities in the trillionths, one would expect to find a large number of matches or near-matches. An analogy can be made to the famous "birthday problem." In its simplest form, the birthday problem assumes that equal numbers of people are born every day of the year. The problem is to determine the minimum number of people in a room such that the odds favor there being at least two of them who were born on the same day of the same month. Focusing solely on the random-match probability of 1/365 for a specified birthday makes it appear that a huge number of people must be in the room for a match to be likely. After all, the chance of a match between two individuals having a given birthday (say, January 1) is (ignoring leap years) a minuscule $1/365 \times 1/365 = 1/133{,}225$. But because the matching birthday can be any

scientists question the utility of the investigative databases for population-genetics research.[455] They observe that these databases contain an unknown number of relatives, that they might contain duplicates, and that the population in the offender databases is highly structured. These complicating factors would need to be considered in testing for an excess of matches or partial matches. Studies of offender databases in Australia and New Zealand that made adjustments for population structure and close relatives showed substantial agreement between the expected and observed numbers of partial matches, at least up to the nine STR loci used for database profiles in those countries at the time of the studies.[456]

The existence of large databases also provides a means of estimating a random-match probability without making any modeling assumptions. For the New Zealand study, even ignoring the possibility of relatives and duplicates, there were only 10 matches out of 59 million comparisons. The empirical estimate of the random-match probability is therefore about 1 in 5.9 million. This is about 10 times larger than the theoretical estimate, but still quite small. As this example indicates, crude but simple empirical estimates from all-pairs comparisons in large databases may well produce random-match probabilities that are larger than the theoretical estimates (as expected when full siblings or other close relatives are in the databases), but the estimated probabilities are likely to remain impressively small.

Investigative Genetic Genealogy (IGG)

Until recently, personal knowledge, oral histories, and documents were the only ways to connect extended family members within and across generations. Since

one of the 365 days in the year and because there are $N(N-1)/2$ ways to have a match, it takes only $N = 23$ people before it is more likely than not that at least two people share a birthday. The birthday problem thus shows that surprising coincidences commonly occur even in relatively small databases. *See, e.g.,* Persi Diaconis & Frederick Mosteller, *Methods for Studying Coincidences*, 84 J. Am. Stat. Ass'n 853 (1989).

455. Bruce Budowle et al., *Partial Matches in Heterogeneous Offender Databases Do Not Call into Question the Validity of Random Match Probability Calculations*, 123 Int'l J. Legal Med. 59 (2009), https://doi.org/10.1007/s00414-008-0239-1.

456. James M. Curran et al., *Empirical Support for the Reliability of DNA Evidence Interpretation in Australia and New Zealand*, 40 Australian J. Forensic Sci. 99, 102–06 (2008), https://doi.org/10.1080/00450610802172230; Bruce S. Weir, *The Rarity of DNA Profiles*, 1 Annals Applied Stat. 358 (2007), https://doi.org/10.1214/07-AOAS128; Bruce S. Weir, *Matching and Partially-Matching DNA Profiles*, 49 J. Forensic Sci. 1009, 1013 (2004); *cf.* Laurence D. Mueller, *Can Simple Population Genetic Models Reconcile Partial Match Frequencies Observed in Large Forensic Databases?*, 87 J. Genetics (India) 101 (2008), https://doi.org/10.1007/s12041-008-0016-4 (maintaining that excess partial matches in an Arizona offender database are not easily reconciled with theoretical expectations). This literature is reviewed in Kaye, *supra* note 454.

2007, direct-to-consumer genetic testing (DTC-GT) has made it possible to find relatives through inherited DNA. Tens of millions of people have undergone such testing for information about their ancestral origin and potential biological relation to other people within the vendors' databases.[457] Law-enforcement officials also have turned to some of these databases to track down the source of crime-scene DNA or to identify human remains indirectly. Like the members of the public interested in finding possible relatives, they try to learn from the companies that assemble these databases if DNA data from potential relatives of the source of the forensic sample seem to be included in the genealogy databases. Constructing parts of the family tree of these relatives has led investigators to the source of the forensic DNA in hundreds of cases.[458] This section explains how this adaptation of personal genetic genealogy is conducted, how it differs from familial searching with law-enforcement databases, how it might be regulated, and how it might implicate the Fourth Amendment. We begin with a brief introduction to the use of DTC-GT databases in genealogy research by members of the public.

How does direct-to-consumer genetic genealogy work?

A DTC-GT company supplies paying customers with mail-in spit kits or cheek swabs. Upon receiving the sample, the company uses certain microarrays (the SNP chips described in the section above titled "Sequence-Specific Probes and Microarrays") to type hundreds of thousands of SNPs. These SNPs are spaced across the 22 autosomal chromosomes, somewhat like individual letters sampled from the volumes of an encyclopedia without the intervening text. Thus, they are not themselves DNA sequences.

Starting in 2009, DTC-GT companies began assembling this information into databases to assist customers seeking to discover possible relatives.[459] A DTC-GT customer can have the company trawl the database with specialized software to identify the chromosomal locations and lengths of blocks of SNPs (haploblocks) that match between the customer's data file and the files

457. James W. Hazel et al., *Direct-to-Consumer Genetic Testing: Prospective Users' Attitudes Toward Information About Ancestry and Biological Relationships*, 16 PLoS One e0260340 (2021), https://doi.org/10.1371/journal.pone.0260340. "This genetic genealogy has enabled thousands of individuals who have lost their biological identity through adoption, abandonment, anonymous gamete donation, misattributed parentage, etc., to regain their genetic heritage." Ellen M. Greytak et al., *Genetic Genealogy for Cold Case and Active Investigations*, 299 Forensic Sci. Int'l 103, 103 (2019), https://doi.org/10.1016/j.forsciint.2019.03.039.

458. Tracey L. Dowdeswell, *Forensic Genetic Genealogy: A Profile of Cases Solved*, 58 Forensic Sci. Int'l: Genetics 102679 (2022), https://doi.org/10.1016/j.fsigen.2022.102679.

459. Daniel Kling et al., *Investigative Genetic Genealogy: Current Methods, Knowledge and Practice*, 52 Forensic Sci. Int'l: Genetics 102474, at 2 (2021), https://doi.org/10.1016/j.fsigen.2021.102474.

in the company's database.[460] Large matching haploblocks—measured in centimorgans[461]—are believed to be "identical by descent" (IBD) and thus indicative of a common recent ancestor.[462] As a rule, the greater the total length of the shared IBD haploblocks, the closer the relationship.[463] The total length thus indicates "the probability that the relationship between the unknown individual and the match falls into each degree of relatedness."[464] But different relationships can correspond to the same range of haploblock sharing.[465] For example, a total of 100 centimorgans could indicate "anywhere from 5th degree to >8th degree, with 6th degree being most likely."[466] Consequently, the customer may have to explore multiple possibilities.[467] Family records, obituaries, and other components of traditional genealogical research will contribute to the proper construction of the family tree.

Someone who wants to maximize the chance of finding relatives can repeat the entire process with another DTC-GT company, but a more efficient strategy is to download the file listing the individual's SNPs and to submit it to other database companies that are willing to accept data files that they did not

460. *See* Greytak et al., *supra* note 457, at 107.

461. On average, a 1 centimorgan (cM) haploblock is about 1 million base pairs long, but a centimorgan is not a fixed physical distance. One cM equates to a 1% probability of recombination within the haploblock when the egg and sperm cells are formed (*see supra* section titled "Sexual Reproduction and the Genome"). Because recombination occurs more often in certain places along the genome than others, the physical distance depends on the location within the genome. Also, because recombination during the formation of egg cells happens more frequently than in the formation of sperm cells, the physical distance also differs between males and females. The randomness of the points at which "crossing over" of chromosomes occurs during meiosis gives rise to the relationship between total haploblock length and types of kinship. For a more complete explanation, see Greytak et al., *supra* note 457, at 104–06.

462. Claire L. Glynn, *Bridging Disciplines to Form a New One: The Emergence of Forensic Genetic Genealogy*, 13 Genes 1381, at 4 (2022), https://doi.org/10.3390/genes13081381; Brenna M. Henn et al., *Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples*, 7 PLoS One e34267 (2012), https://doi.org/10.1371/journal.pone.0034267 (second to ninth degree cousins); Chad Huff et al., *Maximum-Likelihood Estimation of Recent Shared Ancestry (ERSA)*, 21 Genome Rsch. 768 (2011), https://doi.org/10.1101/gr.11597 (third and even fourth degree cousins). However, "the amount of shared DNA can vary greatly for relatives of the same degree," and "~10% of third cousins and ~50% of fourth cousins share no detectable IBD segments." Greytak et al., *supra* note 457, at 106.

463. Ge & Budowle, *supra* note 444, at 434.

464. Greytak et al., *supra* note 457, at 107. A given individual has a first-degree relationship with his or her parents, full siblings, and children; a second-degree relationship with grandparents, aunts and uncles, half siblings, nieces and nephews, and grandchildren; a third-degree relationship with great-grandparents, great-uncles and great-aunts, first cousins, half-nieces and half-nephews, grand-nieces and grand-nephews, and great-grandchildren. For a chart, see *id.* at 105 (fig. 1).

465. *Id.* at 106 (tbl. 3) & 107; Glynn, *supra* note 462, at 5–6.

466. Greytak et al., *supra* note 457, at 106.

467. *Id.* at 107. Additional complications can arise from the history of mating in some groups. *Id.* (noting endogamy and "pedigree collapse").

produce. Indeed, one genealogy resource, known as GEDmatch,[468] does no DNA testing but lets anyone upload a SNP genotype file prepared by any major DTC-GT company. It will then perform an autosomal-haploblock comparison. The individual user receives, at no charge, information on total haploblock sharing with other GEDmatch users and whatever name (which could be an alias) and email address (which could be a one-time-use address) the possible relatives provided. In all these situations, the genealogy enthusiast never sees the raw SNP data file of anyone else.

How does IGG work?

IGG applies the same type of kinship searching to DNA recovered from a crime scene or victim to link an individual to the commission of the crime, or from human remains to identify a missing person.[469] Also called forensic investigative genetic genealogy (FIGG), IGG differs from familial searching (see section above titled "Kinship trawling ('familial searching')"). Familial searching operates within law-enforcement-managed databases and compares STR profiles derived from legally compelled DNA samples. The grasp of this autosomal STR-based kinship trawling is basically limited to first-degree relatives of convicted offenders or arrestees. IGG is another type of "indirect" or "outer-directed" genetic search for individuals who deposited DNA associated with crimes but whose DNA is not in the accessed database. But IGG uses different DNA variants, different software, and different databases. The genealogy databases are populated with SNP DNA profiles that have been voluntarily made available to commercial database vendors for genealogical purposes. Through individually declared consent and privacy options, users of these databases can specify whether they are willing to accept law-enforcement trawling of their information. Unlike familial searching, IGG can support kinship associations beyond first-degree relatives, occasionally looking as far as ninth-degree relatives. As such, policies and legal conclusions reached for familial searching do not necessarily carry over to IGG.

IGG requires three steps to generate leads for police to consider. To begin with, as with all DNA methods, trace-evidence or victim or missing-person

468. A genetic-technology company, Verogen, specializing in forensic-sequencing methods, acquired GEDmatch in 2019, and a larger biotechnology company, Qiagen, acquired Verogen in 2023. Press Release, QIAGEN Completes Acquisition of Verogen, Strengthening Leadership in Human ID/Forensics with NGS Technologies, Jan. 9, 2023, https://perma.cc/8UCR-AEKV.

469. For a comprehensive and detailed review, see Daniel Kling et al., *Investigative Genetic Genealogy: Current Methods, Knowledge and Practice*, 52 Forensic Sci. Int'l: Genetics 102474 (2021), https://doi.org/10.1016/j.fsigen.2021.102474.

DNA must be recovered. This DNA normally will have been analyzed to yield an STR profile suitable for comparison with a known suspect's DNA or for trawling a CODIS database. If conventional STR profiling fails, or if queries within CODIS do not yield any matches or are not feasible because of sample limitations,[470] a more laborious genetic genealogy investigation might proceed. For such an investigation, police agencies often turn to commercial forensic-science service providers, although publicly funded forensic-science laboratories are developing these capabilities.[471]

From the forensic sample to the SNP genotype. The participating laboratory could analyze the DNA in the forensic sample with a SNP chip, as is done for the personal genealogy samples discussed in the preceding section, but microarrays may be less effective with lower quantity, degraded, or contaminated forensic samples.[472] To cope with such samples, massively parallel sequencing methods such as whole-genome or targeted sequencing[473] have been used to identify upward of thousands to millions of SNPs and pull out a set that are either in the DTC-GT microarrays or are compatible with the databases.[474] With a data file from the forensic DNA, the investigation proceeds in essentially the same manner as the usual personal genetic genealogy search.

470. When the quantity of DNA in the forensic sample is very small, and it is doubtful that STR profiling will succeed, rather than waste the DNA, a SNP profile may be generated without first trying STR profiling. In an urgent case, with enough forensic DNA, simultaneously pursuing both STR and haploblock strategies could be attractive.

471. Press Release, Univ. N. Tex. Health Sci. Cntr. at Fort Worth, CHI Becomes Nation's First Public Lab to Earn Accreditation to Perform Forensic Genetic Genealogy (Oct. 12, 2023), https://perma.cc/F9PW-U9PQ.

472. Kling et al., *supra* note 469, § 7.1. *But see* Greytak et al., *supra* note 457, at 103 (reporting success with trace-evidence samples of as little as one nanogram of DNA).

473. *See supra* section titled "Sequencing and SNPs."

474. Ge & Budowle, *supra* note 444, at 433; Kling et al., *supra* note 469, § 7.2. A targeted-sequencing kit that generates a smaller data file of about 10,000 widely spaced SNPs with no known clinically meaningful associations has been developed for use with GEDmatch. June Snedecor et al., *Fast and Accurate Kinship Estimation Using Sparse SNPs in Relatively Large Database Searches*, 61 Forensic Sci. Int'l: Genetics 102769 (2022), https://doi.org/10.1016/j.fsigen.2022.102769; Jessica Watson et al., *Operationalisation of the ForenSeq® Kintelligence Kit for Australian Unidentified and Missing Persons Casework*, 68 Forensic Sci. Int'l: Genetics 102972 (2024), https://doi.org/10.1016/j.fsigen.2023.102972; *see also* Andreas Tillmar et al., *The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications*, 12 Genes (Basel) 1968 (2021), https://doi.org/10.3390/genes12121968 (targeted sequencing kit said to include "3931 autosomal SNPs for extended kinship analysis" and "designed to exclude clinically relevant markers").

From the SNP genotype to possible relatives. The IGG service provider uploads the SNP genotype to the website of a company that maintains a database of SNP profiles for individuals looking for possible relatives (and that is willing to accept data files that it did not generate). Most if not all of these companies have procedures that allow customers to affirmatively opt in or opt out of warrantless forensic IGG trawls.[475] The company's software performs haploblock matching within its database. The matching involves only the length of shared haploblocks;[476] there is no need for the analyst to look at which SNPs are present within these or any other haploblocks in the forensic sample.[477] The database company then informs the investigators of the names or usernames of individuals with enough shared haploblocks to merit attention as possible relatives. The total extent and pattern of the overlap in the SNP genotypes among the "matches" will be available, along with a way to contact the possible relatives. If their identities can be ascertained, traditional genealogy work begins. A genealogist constructs family trees (typically with hundreds of entries for a single third-cousin-level match), usually by scouring public records, newspapers, social media, and so on.[478] The goal is to find a common ancestor for a SNP profile in the genealogy database and the forensic-sample profile. Descendants of that ancestor will include the unknown DNA source, and a combination of the demographic and genetic information may enable a genealogist or crime analyst with genealogy training to whittle down the list of descendants somewhat.[479] The police now have their investigative lead. Often, it is followed by surreptitiously collecting DNA from the individuals in question to ascertain if any of them have the same STR profile as that of the forensic DNA.[480]

Does IGG violate the Fourth Amendment?

A series of Fourth Amendment questions can arise in connection with IGG. At the outset, one might ask whether the steps involved in IGG are a "search or

475. Heather Murphy, *What You're Unwrapping When You Get a DNA Test for Christmas*, N.Y. Times, Dec. 22, 2019.

476. *See supra* section titled "How does direct-to-consumer genetic genealogy work?"

477. Such data normally would exist only in cases in which the SNP profile was derived from more complete genome sequence data.

478. Greytak et al., *supra* note 460, at 108.

479. *Id.* at 107–10; Ge & Budowle, *supra* note 444, at 436 ("Most genealogy investigations focus on third degree or closer relationships because of the substantial burden of searching fourth degree or more matches. An efficient way to improve public record searching is triangulation, which identifies a source by intersecting between two potential families.").

480. Surreptitious DNA sampling is discussed further in the next two sections. In cases in which an STR profile cannot be obtained (*see supra* note 470), SNP-identification profiles can be used for the direct identification.

seizure" of "persons, houses, papers [or] effects."[481] If they are, a further question is whether taking any or all of those steps without a warrant would violate the Amendment's protection against "unreasonable searches and seizures." And finally, if police apply for a warrant, what would be necessary to establish probable cause?

Are parts of IGG a search or seizure? As explained in the previous section, IGG begins with the collection of biological material from a crime scene or a victim's remains and the extraction of DNA. To demonstrate a search or seizure, defendants would have to show either (1) that the forensic-evidence collection and preparation interferes with their right to possess or enjoy their property or personal effects[482] or (2) that simply taking possession of the stain, fluid, or cells and extracting DNA reveals information in which they have a reasonable or legitimate expectation of privacy.[483] Few, if any, defendants have advanced such claims about biological trace evidence.[484]

The next step in the process is the extensive SNP analysis of the forensic sample. Is this laboratory analysis of an unknown person's DNA a search? A number of defendants have challenged the taking—and subsequent STR genotyping—of DNA left behind by suspects in police stations or in public places.[485] By and large, courts have been persuaded that STR profiling as

481. U.S. Const. amend. IV.

482. *See* United States v. Jacobsen, 466 U.S. 109, 113 (1984) ("'seizure' of property occurs when there is some meaningful interference with an individual's possessory interests in that property"). Unlike the collection of DNA directly from the person, acquiring DNA from the crime scene or the remains does not entail a bodily intrusion.

483. Katz v. United States, 389 U.S. 347, 353 (1967); Rakas v. Illinois, 439 U.S. 128, 143 (1978).

484. *Cf.* State v. Hartman, 534 P.3d 423, 431–32 (Wash. Ct. App. 2023) (rejecting defendant's argument that under a provision of the state's constitution guaranteeing that "[n]o person shall be disturbed in his private affairs . . . without authority of law," he "retained a privacy interest" in the DNA extracted from semen on the body of a 12-year-old girl who was found raped and murdered).

485. *E.g.*, United States v. Hicks, No. 2:18-cr-20406-JTF-tmp, 2020 WL 7704556 (W.D. Tenn. May 27, 2020) (noting cases); State v. Burns, 988 N.W.2d 352 (Iowa 2023); *cf.* State v. Athan, 158 P.3d 27 (Wash. 2007) (detectives, posing as a fictitious law firm, sent the defendant a letter inviting him to join a fictitious class-action lawsuit, to obtain DNA from the return envelope to compare to DNA from the crime scene). Commentators have questioned the "abandonment" line of cases. *E.g.*, Elizabeth E. Joh, *Reclaiming "Abandoned" DNA: The Fourth Amendment and Genetic Privacy*, 100 Nw. U. L. Rev. 857–84 (2006); Tracey Maclin, *Government Analysis of Shed DNA Is a Search Under the Fourth Amendment*, 48 Tex. Tech. L. Rev. 287 (2015); Albert E. Scherr, *Genetic Privacy and the Fourth Amendment: Unregulated Surreptitious DNA Harvesting*, 47 Ga. L. Rev. 445 (2013). *But see* Edward J. Imwinkelried & David H. Kaye, *DNA Typing: Emerging or Neglected Issues*, 76 Wash. L. Rev. 413, 437 (2001); David H. Kaye, *Science Fiction and Shed DNA*, 101 Nw. U. L. Rev. Colloquy 62 (2006), https://perma.cc/596A-NKH7 (response to Joh).

conducted in the laboratory solely to produce identifying information from "abandoned" biological material is no more invasive of privacy than examining fingerprints from a crime scene.[486] Singly, or more likely in combination, however, many of the SNPs in the raw data file could serve as markers or predictors of disease.[487] Yet, neither the police nor the laboratories working for them inspect the SNP data files for such markers to make haploblock comparisons.[488] Is this fact sufficient to defeat the claim that the very possession of a list of SNPs in a text file interferes with a reasonable expectation of privacy (even before the identity of the individual whose DNA has been analyzed is known)?[489] If not, would legally enforceable rules against misuse or unauthorized disclosure create a convincing analogy to photographs, fingerprints, or CODIS STR profiles (assuming that the act of comparing such biometric identifiers is not itself a search)? Or is the creation of the SNP data file to find the source of the as-yet-unidentified DNA more like the actual perusal of the text and image files in smartphones[490] or the acquisition of cell-site location information on personal movements over long time periods[491]—both Fourth Amendment searches?

The next step in IGG is the haploblock-sharing trawl. Is the trawl performed by the company that operates the commercial genealogy database a governmental search[492] of the defendant's person or effects? A series of

486. Opinions zeroing in on this issue include *United States v. Davis*, 690 F.3d 226, 243–46 (4th Cir. 2012) (separate search); *Burns*, 988 N.W.2d at 364–65 (no search); *Raynor v. State*, 99 A.3d 753 (Md. 2014) (no search); *Commonwealth v. Arzola*, 26 N.E.3d 185, 190–94 (Mass. 2015) (no search). For academic commentary on STR profiling as a search unto itself, see, for example, Kaye, *supra* note 445; Maclin, *supra* note 485; Murphy, *supra* note 446.

487. To reduce this risk, sequencing kits that generate smaller data files of widely spaced SNPs with no known clinically meaningful associations have been developed. *See supra* note 474.

488. Ellen M. Greytak et al., *Privacy and Genetic Genealogy Data*, 361 Science 857 (2018), https://doi.org/10.1126/science.aav0330; Greytak et al., *supra* note 457, at 106–07 ("Additionally, no sensitive genetic information is disclosed to law enforcement during a genetic genealogy search, as the raw genetic data from GEDmatch users is not accessible . . . GEDmatch simply performs comparisons among samples, returning the lengths and chromosomal locations of shared DNA segments, which are used to determine the approximate relationship between individuals.").

489. Initially, the investigators' SNP data file does not pertain to any known individual, but if IGG is successful and the source of the trace-evidence sample is located, then the data will pertain to that person.

490. Riley v. California, 573 U.S. 373 (2014) (viewing data stored on a cellphone is a search that does not fall within the search-incident-to-arrest exception to the warrant-and-probable-cause requirement).

491. Carpenter v. United States, 585 U.S. 296 (2018).

492. On the question of whether ordinary, non-kinship trawls in law-enforcement STR databases are Fourth Amendment "searches," see *Boroian v. Mueller*, 616 F.3d 60 (1st Cir. 2010) (retrawling after a sentence is served is not a search); *Johnson v. Quander*, 440 F.3d 489, 498–99 (D.C. Cir. 2006) (retrawling after the period of probation is completed is not a search); David H. Kaye, *DNA Database Trawls and the Definition of a Search in* Boroian v. Mueller, 97 Va. L. Rev. in Brief 41 (2011);

subquestions may need to be considered. Did the company knowingly and voluntarily trawl the database to assist the investigation?[493] Or did the company not intend to make its services available to law enforcement,[494] but police submitted SNP data files anyway, pretending to be an ordinary subscriber? Would a customer, having chosen to expose his or her SNP data to trawling on behalf of anyone in the general public, retain a reasonable expectation of privacy in that data?[495] And, even if the trawl were considered a search of the customer's data, could a defendant suppress the evidence on the theory that the

Catherine W. Kimel, Note, *DNA Profiles, Computer Searches, and the Fourth Amendment*, 62 Duke L.J. 933 (2013).

493. FamilyTreeDNA and GEDmatch require law-enforcement agencies to identify themselves as such when submitting a SNP-genotype file. FamilyTreeDNA Law Enforcement Guide, https://perma.cc/PHR5-2B59, last visited, Dec. 16, 2023; GEDmatch, Terms of Service and Privacy Policy, Dec. 30, 2021, https://perma.cc/FJ5G-AL6F.

494. *E.g.*, MyHeritage—Terms and Conditions (June 11, 2023), https://perma.cc/DJP9-D7UZ ("using the DNA Services for law enforcement purposes, forensic examinations, criminal investigations, 'cold case' investigations, identification of unknown deceased people, location of relatives of deceased people using cadaver DNA, and/or all similar purposes, is strictly prohibited, unless a court order is obtained"); GEDmatch, Terms of Service and Privacy Policy (Dec. 30, 2021), https://perma.cc/LF67-BBKD (if you "[o]pt-out . . . [y]our kit WILL NOT be compared with kits submitted by law enforcement to identify perpetrators of violent crimes") (emphasis in original).

495. It could be argued that just because DNA samples are sent to DTC-GT companies for SNP analysis, or because the SNP data are made available to the public for genetic-genealogy trawls, no cognizable search occurs when the government also submits a data file for trawling and receives the results. *Cf., e.g.,* United States v. Miller, 425 U.S. 435, 443 (1976) ("The [customer who maintains a bank account] takes the risk, in revealing his affairs to another, that the information will be conveyed by that person to the Government. . . . [T]he Fourth Amendment does not prohibit the obtaining of information revealed to a third party and conveyed by him to Government authorities, even if the information is revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed."). In *Carpenter v. United States*, 585 U.S. 296 (2018), however, the Court clarified that this "third-party doctrine" depends on the nature of the information compiled by or disclosed to third parties. It held that extensive data on the locations of a person's cell phone relative to cell towers reveals too much about a person's private life to be accessible to police at almost no cost just because the person's movements "might be pertinent to an ongoing investigation." *Id.* at 317. *Carpenter* opens the door for subscribers to claim that the genealogy database trawl is a Fourth Amendment "search" of their data. They can argue that exposing them to the risk of being placed on a family tree known to the government is no less invasive of a legitimate privacy interest than a cellphone customer "effectively [being] tailed every moment of every day for five years." *Id.* at 312. But there is public and scholarly disagreement on the strength of the interests of the subscribers to genetic genealogy services in the secrecy of their biological affiliations and their SNPs. *Compare* James W. Hazel & Christopher Slobogin, *"A World of Difference"? Law Enforcement, Genetic Data, and the Fourth Amendment*, 70 Duke L.J. 705 (2021); Ayesha K. Rasheed, *Personal Genetic Testing and the Fourth Amendment*, 2020 U. Ill. L. Rev. 1249; *and* Natalie Ram, *Genetic Privacy After* Carpenter, 105 Va. L. Rev. 1357 (2019); *with* Teneille R. Brown, *Why We Fear Genetic Informants: Using Genetic Genealogy To Catch Serial Killers*, 21 Colum. Sci. & Tech. L. Rev. 114 (2019); *cf.* Paul Ohm, *The Many Revolutions of* Carpenter, 32 Harv. J.L. & Tech. 357, 385 (2019) (concluding that even for a database of whole-genome

use of the customer's data unreasonably invaded the customer's Fourth Amendment interests?[496]

Warrants and probable cause? If some or all of these steps in IGG are individually or collectively a Fourth Amendment search of the defendant's person, papers, or effects, two broad questions remain: first, whether such searches are constitutionally unreasonable in the absence of a warrant based on probable cause; and, second, if a warrant is sought, what must be shown to establish probable cause. With regard to reasonableness, a few Supreme Court opinions suggest that "our general Fourth Amendment approach [entails] 'examining the totality of the circumstances. . . .'"[497] In IGG cases, parties might argue about the relevance and impact of many circumstances. For example, did anyone actually scrutinize the defendant's SNP profile for medically or socially sensitive information? Are there statutory, administrative, or technical constraints against using the SNP data for such purposes?[498] Is the system designed to avoid unnecessary disclosures of family information, such as misattributed paternity? Is the prospect that law-enforcement authorities will initiate trawls of the database known to the subscribers?[499] These privacy-related factors might affect the magnitude of the individual interests as compared to government's law-enforcement interest.

But directly balancing the two sets of interests to determine the reasonableness of a Fourth Amendment search or seizure is controversial. More often, the Supreme Court has admonished that "searches conducted outside the judicial

sequences "[i]t seems unlikely that a court would require a warrant for DNA evidence held by a private third party based on a straight application of the *Carpenter* factors").

496. The defendant would have to overcome the doctrine that an impermissible intrusion on the security or privacy of another person's papers or effects does not entitle the defendant to suppress the evidence. *See, e.g.,* Rakas v. Illinois, 439 U.S. 128, 133–34 (1978) ("A person who is aggrieved by an illegal search and seizure only through the introduction of damaging evidence secured by a search of a third person's premises or property has not had any of his Fourth Amendment rights infringed."); Minnesota v. Carter, 525 U.S. 83, 91 (1998) (a temporary guest in an apartment does not have standing to challenge the reasonableness of a search of the apartment).

497. United States v. Knights, 534 U.S. 112, 118 (2001). With CODIS STR databases, lower courts using this "totality" approach typically balanced the government's interests in building and operating these intelligence-gathering databases against the individual's legitimate interests in being free from bodily intrusion and maintaining the confidentiality of the STR data and samples. *See supra* section titled "Law-enforcement databases and databanks."

498. *Cf.* Maryland v. King, 569 U.S. 435 (2013) (DNA sampling and CODIS STR profiling of arrestees); Whalen v. Roe, 429 U.S. 589 (1977) (law requiring that a copy of prescriptions for the most dangerous legitimate drugs be sent to the state for computer scanning, storage, and access under secure conditions was a reasonable exercise of state police power).

499. *See supra* section titled "From the SNP genotype to possible relatives" (on policies allowing customers to opt in or opt out of exposure to database trawls for law-enforcement purposes).

process, without prior approval by judge or magistrate, are per se unreasonable under the Fourth Amendment—subject only to a few specifically established and well-delineated exceptions."[500] Under this "warrant preference rule," balancing to determine the reasonableness of a type of search that dispenses with warrants or probable cause still might occur, but only at the level of delineating a new categorical exception to these procedures.[501] For a warrant to issue, a magistrate would need to verify that the police have probable cause. But probable cause to believe what? The Supreme Court has said that probable cause to arrest or search a person exists when the totality of the circumstances indicates "reasonable ground for belief of guilt," and "the belief of guilt [is] particularized with respect to the person to be searched or seized."[502] Genealogy database trawls are not undertaken because of a belief about a particular person. They are initiated in the hope that some individuals unknown to the police have their SNP genotypes in the database and that the company's haploblock-matching software will flag them as possible relatives. Should the probable-cause finding turn on whether the size and likely composition of a company's database indicates a reasonable probability of a useful match occurring in that particular database? Why not all databases combined, since the trawls could be performed in all of them? Or are such warrants too redolent of "the reviled 'general warrants' and 'writs of assistance' of the colonial era, which allowed British officers to rummage through homes in an unrestrained search for evidence of criminal activity"?[503] The nature of probable cause in the context of IGG may need to be explored.

How should IGG be regulated?

After IGG burst into the public consciousness with the arrest in 2018 of a former police officer found to be the elusive Golden State Killer,[504] DTC-GT companies scrambled to develop clearer or different policies,[505] and observers advocated

500. Katz v. United States, 389 U.S. 347, 357 (1967) (notes omitted).

501. David H. Kaye, *A Fourth Amendment Theory for Arrestee DNA and Other Biometric Databases*, 15 U. Pa. J. Const. L. 1095, 1139–40 (2013) (proposing a "biometric exception" that would permit warrantless searches of fingerprint, photograph, and, arguably, CODIS STR databases). For efforts to explain the rationale for the balancing for DNA sampling that the Court performed in *Maryland v. King*, see, for example, Kaye, 104 J. Crim. L. & Criminology 535, *supra* note 431; Maclin, *supra* note 431.

502. Maryland v. Pringle, 540 U.S. 366, 371 (2003) (internal quotation marks and citations omitted).

503. Riley v. California, 573 U.S. 373 (2014).

504. *See, e.g.,* Paige St. John, *The Untold Story of How the Golden State Killer Was Found: A Covert Operation and Private DNA*, L.A. Times, Dec. 8, 2020.

505. Glynn, *supra* note 462, at 10; Murphy, *supra* note 475.

for external rules and oversight of when and how police access the databases.[506] A year and a half later, the Department of Justice settled on an "interim policy" for its agencies, employees, and contractors.[507] This policy

- Mostly limits the use of IGG to unsolved homicides and sex crimes.[508]
- Restricts investigators to "only those GG services that provide explicit notice to their service users and the public that law enforcement may use their service sites."[509]
- Requires that "biological samples and [IGG] profiles [be used] only for law enforcement identification purposes"[510] and not "to determine the sample donor's genetic predisposition for disease or any other medical condition or psychological trait."[511]
- Establishes procedural constraints on surreptitious collection of DNA from certain individuals who are not targets of the investigation.[512]
- Requires that all forensic samples first be tested to generate an STR profile prior to performing IGG, a practice that is not feasible with low input, degraded, and rootless hair samples.

506. Sara H. Katsanis, *Pedigrees and Perpetrators: Uses of DNA and Genealogy in Forensic Investigations*, 21 Ann. Rev. Genomics & Hum. Genetics 535 (2020), https://doi.org/10.1146/annurev-genom-111819-084213; Erin Murphy, *Law and Policy Oversight of Familial Searches in Recreational Genealogy Databases*, 292 Forensic Sci. Int'l e5–e9 (2018), https://doi.org/10.1016/j.forsciint.2018.08.027; Paige St. John, *DNA Genealogical Databases Are a Gold Mine for Police, But with Few Rules and Little Transparency*, L.A. Times, Nov. 24, 2019.

507. U.S. Dep't of Just., Interim Policy Forensic Genetic Genealogical DNA Analysis and Searching, Nov. 1, 2019, https://perma.cc/D6ZF-D6SW. The policy also applies to "any federal agency or any unit of state, local, or tribal government that receives grant award funding from the Department that is used to conduct FGG/FGGS." *Id.* at 2. However, the policy "does not . . . limit the prerogatives, choices, or decisions available to, or made by, the Department in its discretion." *Id.* at 1 n.1.

508. These include missing-person cases where the unidentified human remains are reasonably believed to be those of a suspected homicide victim. *Id.* at 4. However, IGG can be used for other crimes when (1) they are "serious" and the database vendor authorizes "investigative use of its service by law enforcement," *id.* at 4 n.15, or (2) "the circumstances surrounding the criminal act(s) present a substantial and ongoing threat to public safety or national security." *Id.* at 5. Reasonable investigative leads must have been pursued, and a CODIS search must have failed. *Id.*

509. *Id.* at 6.

510. *Id.*

511. *Id.* at 7.

512. The protected individuals are non-suspects "who may have a closer kinship relationship to the donor of the forensic sample than the associated [genetic genealogy] service user" and who have come to light as a result of the genealogical research. *Id.* at 6. In such cases, investigators either must obtain their informed consent or have "reasonable grounds to believe that this request would compromise the integrity of the investigation" and consult prosecutors before covertly collecting their samples. *Id.* In addition, investigators must obtain a search warrant for "a vendor laboratory" to perform SNP typing on the sample. *Id.*

- Requires that only STR profiles be used for confirmatory identity verification against a sample collected from the putative perpetrator or missing person, when SNP-based statistics might more easily be generated with certain samples.
- Creates other constraints and procedural requirements as well.

In sum, the interim policy stakes out a middle ground between unfettered IGG and a total ban. A few states have enacted more restrictive statutes that include judicial oversight.[513] Both the Justice Department policy and the statutes have their critics.[514]

Inferring Traits

Kinship searching (see section titled "Kinship trawling ('familial searching') above)" is one technological response to unsuccessful database searches. Two additional techniques are available: biogeographic ancestry testing and forensic DNA phenotyping. Their purpose is only to focus an investigation by supplying police with some visible or other characteristics that the source of the trace evidence might have. They are not routinely used and are not designed to produce evidence for trial.[515] If a suspect is located with the help of these probabilistic clues, STR profiling (or typing with other loci suitable for individual identification) would be undertaken before trial.[516]

513. *E.g.,* Utah Code § 53-10-403.7 (2023); Natalie Ram et al., *Regulating Forensic Genetic Genealogy: Maryland's New Law Provides a Model for Others*, 373 Science 1444 (2021), https://doi.org/10.1126/science.abj5724.

514. *See* Virginia Hughes, *Two New Laws Restrict Police Use of DNA Search Method: Maryland and Montana Have Passed the Nation's First Laws Limiting Forensic Genealogy, the Method that Found the Golden State Killer*, N.Y. Times, May 31, 2021; Laura Geary, *A Critical Eye Toward Commercial DNA Database Criminal Procedures*, U. Chi. L. Rev. Online, July 8, 2022, https://perma.cc/TG3D-U2GZ (questioning the statutes); Christi Guerrini et al., *State Genetic Privacy Statutes: Good Intentions, Unintended Consequences?*, Bill of Health (Aug. 10, 2023), https://perma.cc/2VGQ-S4L7 (pointing out ambiguities and anomalies); *cf.* Brown, *supra* note 495 (critiquing some of the concerns about IGG); Christi J. Guerrini et al., *Four Misconceptions About Investigative Genetic Genealogy*, 8 J. L. & Biosciences 1 (2021), https://doi.org/10.1093/jlb/lsab001 (urging more nuanced analysis of the ethical and practical issues).

515. *But see* Jonathan Foox et al., *Epigenetic Forensics for Suspect Identification and Age Prediction*, 1 Forensic Genomics 83 (2021), https://doi.org/10.1089/forensic.2021.0005 (experiment with DNA-based age estimation offered to rebut offender's claim that an old DNA sample from a previous case was planted in the newer case for which post-conviction relief was denied).

516. Indeed, police have used the DNA-based information in conducting "mass screens" or "DNA dragnets" of legally consenting residents of a geographic region where serial rapes, murders, or bombings have occurred. *E.g.*, Monroe v. City of Charlottesville, 579 F.3d 380 (4th Cir. 2009). There is a risk that a screening will be restricted on the basis of race or ethnicity to the wrong group. But DNA predictions also could prompt police to change a screen that otherwise would be

Biogeographic ancestry testing (BGAT)

Some DNA sequences are concentrated in some parts of the world. As such, a well-chosen set of loci can be used to infer the geographic area in which an individual's biological ancestors lived. The specificity and accuracy of these inferences depends on several factors. What is the size and scale of the population-genetics databases used to associate genotypes with geographic regions? How clearly do the alleles in the panel of loci cluster according to geography (or self-declared ancestry, if that is used)? How effective is the statistical method for probabilistically mapping genotypes to geography?

These matters are all subjects of ongoing research. As potential ancestry-informative markers (AIMs), the STR loci used for individual identification are far from ideal. They vary greatly among individuals—every population has many alleles at each locus—but it is roughly the same "many" in most populations. For ancestry assignment, various panels of SNPs that are more strongly correlated to geography have been published.[517] Which ancestry-informative SNPs are present in a trace-evidence sample can be determined with most of the methods described above in the section titled "What Are DNA Polymorphisms and How Are They Detected?"[518] Whereas clinical and anthropological research studies often use thousands of SNPs, the panels developed for biological trace evidence, which frequently involves low quantity or degraded DNA, consist of smaller sets of AIMs.[519]

incorrectly concentrated on one group. Imwinkelried & Kaye, *supra* note 485, at 451. On the equal-protection ramifications of racially focused investigations resulting from eyewitness statements, see *Monroe*, *supra*; *Brown v. City of Oneonta*, 221 F.3d 329 (2d Cir. 2000); Imwinkelried & Kaye, *supra* note 485, at 446–51.

517. For a review, see Ozlem Bulbul & Kenneth K. Kidd, *Forensic Ancestry Inference: Data Requirements, Analysis Methods, and Interpretation of Results*, *in* Forensic DNA Analysis: Technological Development and Innovative Applications 225, 227–29 (Elena Pilli & Andrea Berti eds., 2021).

518. *Id.* at 229–31; Elena Pilli et al., *Biogeographical Ancestry, Variable Selection, and PLS-DA Method: A New Panel to Assess Ancestry in Forensic Samples via MPS Technology*, 62 Forensic Sci. Int'l: Genetics 102806 (2023), https://doi.org/10.1016/j.fsigen.2022.102806.

519. Peter Resutik et al., *Comparative Evaluation of the MAPlex, Precision ID Ancestry Panel, and VISAGE Basic Tool for Biogeographical Ancestry Inference*, 64 Forensic Sci. Int'l: Genetics 102850 (2023), https://doi.org/10.1016/j.fsigen.2023.102850 ("More than twenty forensic panels designed to distinguish continental ancestry have been proposed by the forensic community over the past two decades. These generally comprise less than 200 AIMs, mainly consisting of autosomal SNPs, but also including other types of markers such as microhaplotypes (MHs) and insertion-deletion polymorphisms (INDELs). In addition, the inclusion of uniparental markers (Y chromosomal SNPs and mtDNA) in the same panel or as a complementary approach has been advocated as it enables ascertaining information on the maternal and paternal lineages, which may help resolve ancestry of recently-admixed individuals.") (references omitted).

The population-reference databases for developing the ancestry-informative SNPs for crime-scene samples are public research databases.[520] Their limited coverage of populations around the world constrains the inferred origins to very large areas. Basically, forensic BGAT is most reliable for mapping a sample to a major continental group such as African, Western Eurasian, East and South Asian, or Native American. Because significant numbers of SNPs are employed as markers, established multivariate statistical methods are used to make the classifications.[521] Studies of the accuracy of classifications from different providers continue to appear.[522]

Geneticists and anthropologists typically caution that BGA should not be confused with ethnicity and race. Ethnicity reflects a person's social and cultural background. Obviously, these are not determined by DNA. However, a person's ethnicity may be strongly associated with BGA. As a biological classification, the term "race" is considered inappropriate, but self-reported race also is correlated with BGA.[523]

Forensic DNA phenotyping (FDP)

A phenotype is an observable or testable trait—skin color, height, blood sugar level, diseases, physical activity level, etc. Phenotypes result from gene activity and environmental factors. Forensic DNA phenotyping refers to inferring traits from the DNA in a trace-evidence sample to assist in locating the source. As such, it might be clearer to call it DNA trace-evidence phenotyping. Regardless of nomenclature, researchers and practitioners have focused on deducing

520. The 1000 Genomes Project, Human Genome Diversity Project-Center d'Étude du Polymorphisme Humain (HGDP–CEPH), and HapMap Project databases are commonly used. A short description of their coverage, along with that of the ALFRED database of frequency of polymorphisms in human populations, is at Bulbul & Kidd, *supra* note 517, at 231–32.

521. *Id.* at 232–37. Machine-learning methods also are being studied. Eugenio Alladio et al., *Multivariate Statistical Approach and Machine Learning for the Evaluation of Biogeographical Ancestry Inference in the Forensic Field*, 12 Sci. Reports 8974 (2022), https://doi.org/10.1038/s41598-022-12903-0.

522. Muna Al-Asfi et al., *Assessment of the Precision ID Ancestry Panel*, 132 Int'l J. Legal Med. 1581 (2018), https://doi.org/10.1007/s00414-018-1785-9; Lauren Atwood et al., *From Identification to Intelligence: An Assessment of the Suitability of Forensic DNA Phenotyping Service Providers for Use in Australian Law Enforcement Casework*, 11 Frontiers in Genetics 568701 (2021), https://doi.org/10.3389/fgene.2020.568701; Resutik et al., *supra* note 519; Watson et al., *supra* note 487.

523. For discussion of populations and race in forensic genetics, see, for example, Jobling, *supra* note 183; Mark Shriver et al., *Getting the Science and the Ethics Right in Forensic Genetics*, 37 Nature Genetics 449 (2005), https://doi.org/10.1038/ng0505-449; *supra* section titled "Could an Unrelated Person Be the Source?"; *cf.* Alyna T. Khan et al., *Race, Ethnicity, and Genetics Working Group, Recommendations on the Use and Reporting of Race, Ethnicity, and Ancestry in Genetic Research: Experiences from the NHLBI TOPMed Program*, 2 Cell Genomics 100155 (2022), https://doi.org/10.1016/j.xgen.2022.100155.

externally visible characteristics of the kind that might appear in an eyewitness's description.[524]

One such characteristic (or, more precisely, a state that is associated with many aspects of physical appearance) is biological sex. A locus for sex typing has been part of standard STR profiling for over 25 years.[525] Eye color is harder to infer for all people because many genes are involved, but only six are currently used in FDP tests. Some classifications (such as brown eye color versus blue) often can be made,[526] and, when they are possible, they seem to be generally accurate.[527] Classifications of hair and skin pigmentation also are included in some FDP-typing systems whose validity has been studied in a number of populations.[528] Predicting other externally visible traits such as height is even harder because these quantitative traits are determined by large numbers of genes, many of which remain unknown. These traits are also influenced by environmental factors such as nutrition, which cannot be predicted from DNA. Although some police agencies have turned to bioinformatics companies that construct pictures of faces from DNA, this procedure has been criticized as unvalidated, not peer reviewed, and beyond existing knowledge.[529]

In addition to FDP for characteristics of eyes, hair, and skin, estimation of chronological age is possible.[530] Mitochondrial DNA changes with age (and not

524. Scientific literature on FDP is referenced in John M. Butler, *Recent Advances in Forensic Biology and Forensic DNA Typing: INTERPOL Review 2019–2022*, 6 Forensic Sci. Int'l: Synergy 100311 (2023), https://doi.org/10.1016/j.fsisyn.2022.100311.

525. *See supra* note 48.

526. Eye color depends on how much of a dark brownish-black pigment called eumelanin is in the iris. Eumelanin absorbs light, so individuals who have a lot of eumelanin in their irises have eye colors that appear dark, such as brown. Several genes responsible for controlling eumelanin production in the eye are known. One of the most important (OCA2) is located on chromosome 15. Individuals with two copies of the G nucleotide at a locus near OCA2 produce less eumelanin and are more likely to have blue eyes. Individuals with one or two copies of an A nucleotide at the same SNP locus produce more eumelanin and are more likely to have brown eyes. Intermediate eye colors such as green and hazel are still very hard to predict.

527. *E.g.*, Ersilia Paparazzo et al., *A New Approach to Broaden the Range of Eye Colour Identifiable by IrisPlex in DNA Phenotyping*, 12 Sci. Reports 12803 (2022), https://doi.org/10.1038/s41598-022-17208-w (describing the system as "well validated").

528. *See* European Forensic Genetics Network of Excellence, Making Sense of Forensic Genetics 31 (2017), https://perma.cc/4RSH-5EQM ("Currently, eye colour, hair colour and skin colour can be predicted reliably and with practically useful accuracy from crime scene DNA, but not yet any other externally visible characteristic.") (quoting Manfred Kayser); *cf.* Matteo Fabbri et al., *Application of Forensic DNA Phenotyping for Prediction of Eye, Hair and Skin Colour in Highly Decomposed Bodies*, 11 Healthcare 647 (2023), https://doi.org10.3390/healthcare11050647.

529. *Id.*; Jobling, *supra* note 183.

530. Gerontology researchers often distinguish between chronological and biological age (also known as physiological or functional age). Biological aging occurs as damage to various cells and tissues in the body accumulates. Biological age reflects chronological age, genetics (for example, how quickly antioxidant defenses kick in), lifestyle, nutrition, and diseases and other

for the better),[531] but the association of these modifications with age is weak. In nuclear DNA, sequences near the ends of the chromosomes shorten with every cell division. Estimates based on this effect are not very precise (typical errors are around ±20 years, perhaps less when using certain blood cells). Better results come from changes to DNA that do not alter the sequence of nucleotides.[532] Such "epigenetic" changes occur in connection with gene expression. Messenger RNA (mRNA) participates in this gene expression, and other molecules attach to the DNA in the process that turns gene expression on and off.[533] In particular, a methyl group (one carbon atom and three hydrogen atoms) on top of a gene turns it off (or sometimes on). The extent and pattern of this methylation across the genome changes as an individual ages. This phenomenon enables DNA methylation and mRNA to serve as age markers. Analytical procedures for markers in large trace-evidence samples seem to give estimated ages that usually are accurate within about ±3 or 4 years; massively parallel sequencing for methylation in samples containing smaller amounts of DNA is in development[534] and has been used in one prominent criminal case.[535]

FDP has prompted criticism from some legal commentators and bioethicists.[536] Proponents of the methods maintain that FDP merely reveals information for which a person leaving a trace-evidence sample could have no plausible expectation of privacy, since visible traits are exposed to the public anyway. Beyond phenotypes like pigmentation, "lifestyle factors, such as smoking, alcohol intake, drug abuse, diet, physical exercise and educational attainment have been reported to impact our epigenome" and to merit research.[537] But some commentators regard "lifestyle and environmental characteristics" as included in "the right to privacy."[538] Other writers note that "if FDP techniques

conditions. Biological age is more closely connected to appearance, but chronological age can be more directly useful in investigations.

531. A higher load of point heteroplasmies and an increasing number of large-scale deletions occur.

532. Vidaki & Kayser, *supra* note 398.

533. *See supra* section titled "Genes and Gene Products."

534. Walther Parson, *Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review*, 64 Gerontology 326 (2018), https://doi.org/10.1159/000486239; Vidaki & Kayser, *supra* note 398.

535. State v. Avery, 970 N.W.2d 564 (Wis. Ct. App. 2021) (table); Foox et al., *supra* note 515.

536. Gabrielle Samuel & Barbara Prainsack, Societal, Ethical, and Regulatory Dimensions of Forensic DNA Phenotyping (2019), https://perma.cc/XT53-H35S (surveying literature and summarizing interviews); Pilar N. Ossorio, *About Face: Forensic Genetic Testing for Race and Visible Traits*, 34 J. L., Med. & Ethics 277 (2006), https://doi.org/10.1111/j.1748-720X.2006.00033.x; Victor Toom et al., *Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively*, 22 Forensic Sci. Int'l: Genetics e1 (2016), https://doi.org/10.1016/j.fsigen.2016.01.010.

537. Vidaki & Kayser, *supra* note 398.

538. Matthias Wienroth et al., *Ethics as Lived Practice. Anticipatory Capacity and Ethical Decision-Making in Forensic Genetics*, 12 Genes 1868 (2021), https://doi.org/10.3390/genes12121868.

were extended to [certain] non-visible characteristics, such as genetic risk of disease, they could reveal personally sensitive, private information, to whoever is doing the testing—which could be of interest to medical insurance companies or certain employers."[539] In addition, there are questions of premature use of methods that have not attained adequate validation and precision, and of effectively communicating uncertainties. Also, there is concern that police investigations informed by surmises about skin color will "be seen as reinforcing racial stereotyping and perceptions of unequal treatment amongst minority communities."[540]

Whether trace-evidence samples can be used for phenotyping has not been litigated. For samples collected for storage in law-enforcement databanks, most legislation establishing and funding law-enforcement DNA databases simply uses the broad term "identification" to denote the purpose behind collecting and storing samples, presumably from convicted offenders or arrestees rather than from the trace evidence. However, a few states exclude testing these samples for "physical characteristics, traits or predispositions for disease."[541]

539. European Forensic Genetics Network of Excellence, *supra* note 528. An outpouring of legislation forbids the use of genetic-test results in health insurance and employment. *See generally* Jean-Christophe Bélisle-Pipon et al., *Genetic Testing Insurance Discrimination and Medical Research: What the United States Can Learn from Peer Countries*, 25 Nature Med. 1198 (2019), https://doi.org/10.1038/s41591-019-0534-z; Mark A. Rothstein, *Time to End the Use of Genetic Test Results in Life Insurance Underwriting*, 46 J.L., Med. & Ethics 794 (2018), https://doi.org/10.1177/1073110518804243 ("By the end of the [1990s], 48 states had enacted laws prohibiting genetic discrimination in health insurance and 35 states prohibited genetic discrimination in employment. In 2008, Congress enacted the Genetic Information Nondiscrimination Act (GINA), which outlawed discrimination based on genetic information in health insurance and employment.") (notes omitted).

540. Weinroth et al., *supra* note 538, at 32 (quoting Robin Williams).

541. Indiana Code § 10-13-6-16 ("The information contained in the Indiana DNA data base may not be collected or stored to obtain information about human physical traits or predisposition for disease."); Wyo. Stat. § 7-19-404(c) (2022) ("Only DNA records which directly relate to the identification characteristics of individuals shall be collected and stored in the state DNA database. The information contained in the state DNA database shall not be collected or stored for the purpose of obtaining information about physical characteristics, traits or predisposition for disease."); *cf.* R.I. Gen. L. § 12–1.5-10(5) (2022) ("DNA samples and DNA records collected under this chapter shall never be used under the provisions of this chapter for the purpose of obtaining information about physical characteristics, traits or predispositions for disease."). On the desirability of limiting the loci that may be typed from offender samples in law-enforcement databanks, compare Mark A. Rothstein & Sandra Carnahan, *Legal and Policy Issues in Expanding the Scope of Law Enforcement DNA Data Banks*, 67 Brook. L. Rev. 127 (2001), with David H. Kaye, *Two Fallacies about DNA Data Banks for Law Enforcement*, 67 Brook. L. Rev. 179 (2001). A few European countries have laws that affect phenotyping with trace-evidence samples. Henrik Westermark et al., Swiss Inst. Comparative L., The Regulation of the Use of DNA in Law Enforcement, Aug. 28, 2020, at 10, https://perma.cc/B3K6-DK3T.

Glossary of Terms

adenine (A). One of the four bases, or nucleotides, that make up the DNA molecule. Adenine binds only to thymine. See nucleotide.

affinal method. A method for computing the single-locus profile probabilities for a theoretical subpopulation by adjusting the single-locus profile probability, calculated with the product rule from the mixed population database, by the amount of heterogeneity across subpopulations. The model is appropriate even if there is no database available for a particular subpopulation, and the formula always gives more conservative probabilities than the product rule applied to the same database.

allele. In classical genetics, an allele is one of several alternative forms of a gene. A biallelic gene has two variants; others have more. Alleles are inherited separately from each parent, and for a given gene an individual may have two different alleles (heterozygosity) or the same allele (homozygosity). In DNA analysis, the term is applied to any DNA region (even if it is not a gene) used for analysis.

allelic ladder. A mixture of all the common alleles at a given locus. Periodically producing electropherograms of the allelic ladder aids in designating the alleles detected in an unknown sample. The positions of the peaks for the unknown can be compared to the positions in a ladder electropherogram produced near the time when the unknown was analyzed. Peaks that do not match up with the ladder require further analysis.

amplification. Increasing the number of copies of a DNA region, usually by PCR.

amplified fragment length polymorphism (AMP-FLP). A DNA identification technique that uses PCR-amplified DNA fragments of varying lengths. The DS180 locus is a VNTR whose alleles can be detected with this technique.

ancestry informative markers (AIMs). Polymorphisms for a particular DNA sequence that appear in substantially different frequencies among populations from different geographical regions of the world. AIMs are used to infer the geographical origins of the ancestors of an individual, typically by continent or subcontinent. Panels of SNPs are usually used, as they vary more in their frequency across major populations than do the STRs used in individual identification.

antibody. A protein (immunoglobulin) molecule, produced by the immune system, that recognizes a particular foreign antigen and binds to it; if the antigen is on the surface of a cell, this binding leads to cell aggregation and subsequent destruction.

antigen. A molecule (typically found on the surface of a cell) whose shape triggers the production of antibodies that will bind to the antigen.

autoradiograph (autoradiogram, autorad). In RFLP analysis, the X-ray film (or print) showing the positions of radioactively marked fragments (bands) of DNA, indicating how far these fragments have migrated, and hence their molecular weights.

autosome. A chromosome other than the X and Y sex chromosomes.

band. See autoradiograph.

band shift. Movement of DNA fragments in one lane of a gel at a different rate than fragments of an identical length in another lane, resulting in the same pattern "shifted" up or down relative to the comparison lane. Band shift does not necessarily occur at the same rate in all portions of the gel.

base pair (bp). Two complementary nucleotides bonded together at the matching bases (A and T or C and G) along the double helix "backbone" of the DNA molecule. The length of a DNA fragment often is measured in numbers of base pairs (1 kilobase (kb) = 1,000 bp); base-pair numbers also are used to describe the location of an allele on the DNA strand.

Bayes' factor. A ratio that indicates how much the data change a person's prior odds according to Bayes' rule. In simple situations, the Bayes' factor has the same value as the likelihood ratio.

Bayes' rule. A formula that relates certain conditional probabilities. It can be used to describe the impact of new data on the probability that a hypothesis is true. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

bin, fixed. In VNTR profiling, a bin is a range of base pairs (DNA fragment lengths). When a database is divided into fixed bins, the proportion of bands within each bin is determined and the relevant proportions are used in estimating the profile frequency.

bin, floating. In VNTR profiling, a bin is a range of base pairs (DNA fragment lengths). In a floating bin method of estimating a profile frequency, the bin is centered on the base-pair length of the allele in question, and the width of the bin can be defined by the laboratory's matching rule (e.g., ±5% of band size).

binning. Grouping VNTR alleles into sets of similar sizes because the alleles' lengths are too similar to differentiate.

biogeographic ancestry testing (BGA, BGAT). Inferring the geographic region where a person's ancestors lived from DNA variants. Clinical and anthropological research studies often use thousands of SNPs for this purpose. Much smaller panels of these ancestry-informative markers have been developed for biological trace evidence.

blind proficiency test. See proficiency test.

capillary electrophoresis. A method for separating DNA fragments (including STRs) according to their lengths. A long, narrow tube is filled with an entangled polymer or comparable sieving medium, and an electric field is applied to pull DNA fragments placed at one end of the tube through the medium. The procedure is faster and uses smaller samples than gel electrophoresis, and it can be automated.

ceiling method. A procedure for setting a minimum DNA profile frequency proposed in 1992 by a committee of the National Academy of Sciences but not recommended by a second committee in 1996. One hundred persons from each of 15 to 20 genetically homogeneous populations spanning the range of racial groups in the United States are sampled. For each allele, the higher frequency among the groups sampled (or 5%, whichever is larger) is used in calculating the profile frequency. The committees called this procedure, intended to overstate the population frequency of a profile, the ceiling principle. Compare interim ceiling principle.

centimorgan (cM). A unit of measure for the frequency of genetic recombination. One centimorgan is equal to a 1% chance that two markers on a chromosome will become separated from one another owing to a recombination event during meiosis (which occurs during the formation of egg and sperm cells).

chip. A miniaturized system for genetic analysis. One such microfluidic chip mimics capillary electrophoresis and related manipulations. DNA fragments, pulled by small voltages, move through tiny channels etched into a small block of glass, silicon, quartz, or plastic. This system is useful in analyzing STRs. Another technique places a dense array of oligonucleotide probes on a solid surface. Such hybridization microarrays are useful in identifying SNPs and in sequencing mitochondrial DNA.

chromosome. A rodlike structure composed of DNA, RNA, and proteins. Most normal human cells contain 46 chromosomes, 22 autosomes and a sex chromosome (X) inherited from the mother, and another 22 autosomes and a sex chromosome (either X or Y) inherited from the father. The genes are located along the chromosomes. See also homologous chromosomes.

coding and noncoding DNA. The sequence of base pairs in a gene that correspond to the building blocks (amino acids) of a protein is called coding DNA. (A sequence of three base pairs, called a codon, specifies a particular one of the 20 possible amino acids in the protein. The mapping of a set of three nucleotide bases to a particular amino acid is the genetic code. The cell makes the protein through intermediate steps involving coding RNA transcripts.) About 1.5% of the human genome codes for the amino-acid

sequences. Another 23.5% of the genome is classified as genetic sequence but does not encode proteins. This portion of the noncoding DNA is involved in regulating the activity of genes. It includes promoters, enhancers, and repressors. Other gene-related DNA consists of introns (that interrupt the coding sequences, called exons, in genes and that are edited out of the RNA transcript for the protein), pseudogenes (evolutionary remnants of once-functional genes), and gene fragments. The remaining extragenic DNA (about 75% of the genome) also is noncoding.

CODIS (combined DNA index system). A collection of databases on STR and other loci of DNA from convicted felons, arrestees, missing persons, and crime scenes maintained by the FBI.

complementary sequence. The sequence of nucleotides on one strand of DNA that corresponds to the sequence on the other strand. For example, if one sequence is CTGAA, the complementary bases are GACTT.

control region. See D-loop.

cytoplasm. A jelly-like material (80% water) that fills the cell.

cytosine (C). One of the four bases, or nucleotides, that make up the DNA double helix. Cytosine binds only to guanine. See nucleotide.

databank. A collection of DNA samples; a sample repository.

database. A computer-searchable collection of DNA profiles.

degradation. The breaking down of DNA by chemical or physical means.

denature, denaturation. The process of splitting, as by heating, two complementary strands of the DNA double helix into single strands in preparation for hybridization with biological probes.

deoxyribonucleic acid (DNA). The molecule that contains genetic information. DNA is composed of nucleotide building blocks, each containing a base (A, C, G, or T), a phosphate, and a sugar. These nucleotides are linked together in a double helix—two strands of DNA molecules paired up at complementary bases (A with T, C with G). See adenine, cytosine, guanine, thymine.

diploid number. See haploid number.

D-loop. A portion of the mitochrondrial genome known as the control region or displacement loop that is instrumental in the regulation and initiation of mtDNA gene products. Two short "hypervariable" regions within the D-loop that do not appear to be functional are used in identity or kinship testing.

DNA polymerase. The enzyme that catalyzes the synthesis of double-stranded DNA.

DNA probe. See probe.

DNA profile. A list of the alleles at each locus. See genotype.

DNA sequence. The ordered list of base pairs in a duplex DNA molecule or of bases in a single strand.

DQ. The antigen that is the product of the DQA gene. See DQA, human leukocyte antigen.

DQA. The gene that codes for a particular class of human leukocyte antigen (HLA). This gene has been sequenced completely and can be used for forensic typing. See human leukocyte antigen.

electropherogram. The PCR products separated by capillary electrophoresis can be labeled with a dye that glows at a given wavelength in response to light shined on it. As the tagged fragments pass the light source, an electronic camera records the intensity of the fluorescence. Plotting the intensity as a function of time produces a series of peaks, with the shorter fragments producing peaks sooner. The intensity is measured in relative fluorescent units and is proportional to the number of glowing fragments passing by the detector. The graph of the intensity over time is an electropherogram.

electrophoresis. See capillary electrophoresis, gel electrophoresis.

endonuclease. An enzyme that cleaves the phosphodiester bond within a nucleotide chain.

environmental insult. Exposure of DNA to external agents such as heat, moisture, and ultraviolet radiation, or chemical or bacterial agents. Such exposure can interfere with the enzymes used in the testing process or otherwise make DNA difficult to analyze.

enzyme. A protein that catalyzes a reaction (speeds it up without being consumed).

epigenetic. Heritable changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence. Epigenetic marks are molecules attached to DNA that can determine whether genes are active and used by the cell.

ethidium bromide. A molecule that can intercalate into DNA double helices when the helix is under torsional stress. Used to identify the presence of DNA in a sample by its fluorescence under ultraviolet light.

exon. See coding and noncoding DNA.

fallacy of the transposed conditional. See transposition fallacy.

false match. Two samples of DNA that have different profiles could be declared to match if, instead of measuring the distinct DNA in each sample, there is an error in handling or preparing samples such that the DNA from a single sample is analyzed twice. The resulting match, which does not reflect the true profiles of the DNA from each sample, is a false match. Some people use "false match"

more broadly, to include cases in which the true profiles of each sample are the same, but the samples come from different individuals. Compare true match. See also match, random match.

forensic DNA phenotyping (FDP). Inferring observable traits of individuals whose DNA is in a trace-evidence sample, or unknown deceased (missing) persons, directly from biological materials found at the scene. FDP is usually done with panels of SNPs in genes that influence external visible traits such as eye and hair color, but the term may include biogeographic ancestry testing.

forensic genetic genealogy (FGG). Another term for investigative genetic genealogy.

gel, agarose. A semisolid medium used to separate molecules by electrophoresis.

gel electrophoresis. In RFLP analysis, the process of sorting DNA fragments by size by applying an electric current to a gel. The different-sized fragments move at different rates through the gel.

gene. A set of nucleotide base pairs on a chromosome that contains the "instructions" for the product that controls some cellular function such as making an enzyme. The gene is the fundamental unit of heredity; each simple gene "codes" for a specific biological characteristic.

gene frequency. The relative frequency (proportion) of an allele in a population.

genetic drift. Random fluctuation in a population's allele frequencies from generation to generation.

genetic genealogy. Determining family history (biological relationships between or among individuals) using DNA test results in combination with traditional genealogical methods. Direct-to-consumer genetic testing companies analyze DNA samples at hundreds of thousands of single-nucleotide polymorphisms spread across the genome. The locations and lengths of shared haploblocks are used to ascertain the degree of relatedness, and documentary and other research leads to the reconstruction of relevant parts of a family tree.

genetics. The study of the patterns, processes, and mechanisms of inheritance of biological characteristics.

genome. The complete genetic makeup of an organism, including roughly 23,000 genes and many other DNA sequences in humans. The haploid human genome comprises over three billion nucleotide base pairs.

genotype. The particular forms (alleles) of a set of genes possessed by an organism (as distinguished from phenotype, which refers to how the genotype expresses itself, as in physical appearance). In DNA analysis, the term is applied to the variations within one or all DNA regions (whether or not they constitute genes) that are analyzed.

genotype, multilocus. The alleles that an organism possesses at several sites in its genome.

genotype, single locus. The alleles that an organism possesses at a particular site in its genome.

guanine (G). One of the four bases, or nucleotides, that make up the DNA double helix. Guanine binds only to cytosine. See nucleotide.

haploblock (haplotype block). Several neighboring, tightly linked SNPs inherited together as a block. A haploblock has a higher discrimination power than any individual SNP within the block.

haploid number. Human sex cells (egg and sperm) contain 23 chromosomes each. This is the haploid number. When a sperm cell fertilizes an egg cell, the number of chromosomes doubles to 46. This is the diploid number.

haplotype. A specific combination of linked alleles at several loci on one chromosome. "Linked" means that the alleles tend to be inherited together because they are close to each other on the chromosome.

Hardy-Weinberg equilibrium. A condition in which the allele frequencies within a large, random, intrabreeding population are unrelated to patterns of mating. In this condition, the occurrence of alleles from each parent will be independent and have a joint frequency estimated by the product rule. See independence. Compare linkage disequilibrium.

heteroplasmy, heteroplasty. The condition in which some copies of mitochondrial DNA sequences in the same individual have different base pairs or lengths.

heterozygous. Having a different allele at a given locus on each of a pair of homologous chromosomes. See allele. Compare homozygous.

homologous chromosomes. The 44 autosomes (nonsex chromosomes) in the normal human genome are in homologous pairs (one from each parent) that share an identical set of genes, but may have different alleles at the same loci.

homozygous. Having the same allele at a given locus on each of a pair of homologous chromosomes. See allele. Compare heterozygous.

human leukocyte antigen (HLA). Antigen (foreign body that stimulates an immune system response) located on the surface of most cells (excluding red blood cells and sperm cells). HLAs differ among individuals and are associated closely with transplant rejection. See DQA.

hybridization. Pairing up of complementary strands of DNA from different sources at the matching base-pair sites. For example, a primer with the sequence AGGTCT would bond with the complementary sequence TCCAGA on a DNA fragment.

identical by descent (IBD). Haplotypes that are identical and inherited from a common ancestor. IBD blocks are broken up by recombination during meiosis, so their expected length depends on the number of generations since the common ancestor at the locus. If the common ancestor lived a great many generations ago (ancient IBD), the individuals share very short tracts of genetic material. At the other extreme, in families, individuals who have IBD typically share very long tracts (greater than 10 centimorgans), and IBD is only defined with respect to documented common ancestry. Recent IBD is IBD between individuals of possibly undocumented relationship, and it results from common ancestry within approximately the past 30 generations.

INDEL. A general term that may refer to insertion or deletion of nucleotides in genomic DNA. Compare single nucleotide polymorphism.

independence. Two events are said to be independent if one is neither more nor less likely to occur when the other does.

investigative genetic genealogy (IGG). The application of genetic genealogy to generate a list of individuals whose DNA might match that of a trace-evidence or missing-person DNA sample. Also called forensic genetic genealogy (FGG) or forensic investigative genetic genealogy (FIGG).

interim ceiling principle. A procedure proposed in 1992 (and no longer used) by a committee of the National Academy of Sciences for setting a minimum DNA profile frequency. For each allele, the highest frequency (adjusted upward for sampling error) found in any major racial group (or 10%, whichever is higher) is used in product-rule calculations. Compare ceiling principle.

intron. See coding and noncoding DNA.

kilobase (kb). A measure of DNA length (1,000 bases).

kinship analysis. Evaluation of biological relatedness between individuals based on DNA. Kinship analysis is used in parentage testing (civil or criminal), disaster victim identification, missing persons identification, familial searching, genetic genealogy, and immigration control.

kinship index. A likelihood ratio used for kinship analysis. See paternity index.

likelihood ratio (LR). A measure of the support that an observation provides for one hypothesis as opposed to an alternative hypothesis. The likelihood ratio is computed by dividing the conditional probability of the observation given that one hypothesis is true by the conditional probability of the observation given the alternative hypothesis. For example, the likelihood ratio for the hypothesis that two DNA samples with the same correctly ascertained STR profile originated from the same individual (as opposed to originating from two unrelated individuals) is the reciprocal of the

random-match probability. Legal scholars have introduced the likelihood ratio as a measure of the probative value of evidence. Evidence that is 100 times more probable to be observed when one hypothesis is true as opposed to another has more probative value than evidence that is only twice as probable.

lineage marker. DNA features that are inherited exclusively from one parent. Lineage markers are found on sex chromosomes and in mitochondrial DNA.

linkage. The inheritance together of two or more genes or loci on the same chromosome.

linkage equilibrium. A condition in which the occurrence of alleles at different loci is independent.

locus. A location in the genome, that is, a position on a chromosome where a gene, marker, or other structure begins.

Markov chain Monte Carlo approximation. A computer simulation method that can be used to get an approximate solution for the integration that must be performed to apply Bayes' rule to a problem with many random variables. It is used in some computer programs that give ratios of probabilities of the data given the different sets of genotypes that might be present in a DNA sample.

massively parallel sequencing (MPS). Any of several high-throughput methods for determining nearly all or just parts of the nucleotide sequence of an individual's genome. Multiple DNA segments are analyzed simultaneously. There are several instruments with different designs to accomplish this. Also called next-generation sequencing (NGS). See reads.

match. The presence of the same allele or alleles in two samples. Two DNA profiles are declared to match when they are indistinguishable in genotype. For loci such as STRs, with discrete alleles, two samples match when they display the same set of alleles. For RFLP testing of VNTRs, two samples match when the pattern of the bands is similar and the positions of the corresponding bands at each locus fall within a preset distance. See match window, false match, true match.

match window. If two RFLP bands lie within a preset distance (called the match window) that reflects normal measurement error, the bands can be declared to match.

meiosis. A special type of cell division of germ cells that produces sperm or egg cells with only one copy of each chromosome (a haploid genome). In humans, body (somatic) cells are diploid, containing two sets of these chromosomes (one from each parent via the mother's egg cell and the father's sperm cell). Compare mitosis.

messenger RNA (mRNA). A type of single-stranded RNA that carries information from the DNA in a cell's nucleus to the cell's watery interior, where the protein-making machinery reads the mRNA sequence and translates it into the amino acid in a growing protein chain. See coding and noncoding DNA.

methylation. A chemical reaction in the body in which a small molecule called a methyl group gets added to DNA, proteins, or other molecules. The addition of methyl groups can affect how some molecules act in the body. Methylation turns gene expression on or off. See epigenetic.

microarray. See chip.

microfluidics. The technology of manufacturing microminiaturized devices containing chambers and tunnels through which fluids flow or are confined.

microhaplotype. Microhaplotype loci (microhaps) are markers less than 300 nucleotides long that display multiple allelic combinations (polymorphism).

microsatellite. Another term for an STR.

minisatellite. Another term for a VNTR.

mitochondria. A structure (organelle) within nucleated (eukaryotic) cells that is the site of the energy-producing reactions within the cell. Mitochondria contain their own DNA (often abbreviated as mtDNA), which is inherited only from mother to child.

mitosis. The process by which a cell replicates its chromosomes and then segregates them, producing two identical nuclei in preparation for cell division. Mitosis is generally followed by equal division of the cell's content into two daughter cells that have identical genomes. The copying of DNA from a human somatic (body) cell that splits to produce a daughter cell occurs in mitosis. Compare meiosis.

molecular weight. The weight in grams of 1 mole (approximately 6.02×10^{23} molecules) of a pure, molecular substance.

monomorphic. A gene or DNA characteristic that is almost always found in only one form in a population. Compare polymorphism.

multilocus probe. A probe that marks multiple sites (loci). RFLP analysis using a multilocus probe will yield an autorad showing a striped pattern of thirty or more bands. Such probes are no longer used in forensic applications.

multilocus profile. See profile.

multiplexing. Typing several loci simultaneously.

mutation. The process that produces a gene or chromosome set differing from the type already in the population; the gene or chromosome set that results from such a process.

nanogram (ng). A billionth of a gram.

next-generation sequencing (NGS). Any implementation of massively parallel sequencing. There are several generations of NGS. See reads.

nucleic acid. RNA or DNA.

nucleotide. A unit of DNA consisting of a base (A, C, G, or T) and attached to a phosphate and a sugar group; the basic building block of nucleic acids. See deoxyribonucleic acid.

nucleus. The membrane-covered portion of a eukaryotic cell containing most of the DNA and found within the cell's watery interior (the cytoplasm).

oligonucleotide. A synthetic polymer made up of fewer than 100 nucleotides; used as a primer or a probe in PCR. See primer.

paternity index. A number (technically, a likelihood ratio) that indicates the support that the paternity test results lend to the hypothesis that the alleged father is the biological father as opposed to the hypothesis that another man selected at random is the biological father. Assuming that the observed DNA genotypes (or phenotypes) correctly represent those of the mother, child, and alleged father tested, the number can be computed as the ratio of the probability of the genotypes (or phenotypes) under the first hypothesis to the probability under the second hypothesis. The paternity index is one type of kinship index; the siblingship index is another.

pH. A measure of the acidity of a solution.

phenotype. An observable trait, such as height, eye color, or blood group, resulting from a genotype and environmental factors.

picogram. A trillionth of a gram. A human cell contains about six picograms of DNA.

point mutation. See SNP.

polymarker. A commercially marketed set of PCR-based DNA tests for certain protein polymorphisms.

polymerase chain reaction (PCR). A process that mimics DNA's own replication processes to make up to millions of copies of short strands of genetic material in a few hours.

polymorphism. The presence of several forms of a gene or DNA characteristic in a population. Mutations give rise to polymorphisms. In a genome sequencing project, single nucleotide polymorphisms (SNPs) and DNA mutations are defined as DNA variants detectable in more than 1% or in less than 1% of the population, respectively. Compare allele.

population genetics. The study of the genetic composition of groups of individuals.

population structure. When a population is divided into subgroups that do not mix freely, that population is said to have structure. Significant structure can lead to allele frequencies being different in the subpopulations.

primer. An oligonucleotide that attaches to one end of a DNA fragment and provides a point for more complementary nucleotides to attach and replicate the DNA strand. See oligonucleotide.

probabilistic genotyping. Electropherograms display patterns of peaks that could result from different genotypes. Each possible genotype has its own probability of giving rise to the observed pattern (the data). In traditional binary matching, every genotype is said to match with a probability of 0 or 1, and a likelihood ratio for the single matching genotype (or single set of genotypes in a mixed profile) is assigned. Probabilistic genotyping allows the possible genotypes to give rise to the data with probabilities between 0 and 1. It uses statistical modeling to compute these probabilities. Likelihood ratios are reported for the possible genotypes (or combinations of genotypes that could be in a mixed profile).

probe. In forensics, a short segment of DNA used to detect certain alleles. The probe hybridizes, or matches up, to a specific complementary sequence. Probes allow visualization of the hybridized DNA, either by a radioactive tag (usually used for RFLP analyses) or a biochemical tag (usually used for PCR-based analyses).

product rule. When alleles occur independently at each locus (Hardy-Weinberg equilibrium) and across loci (linkage equilibrium), the proportion of the population with a given genotype is the product of the proportion of each allele at each locus, times factors of two for heterozygous loci.

proficiency test. A test administered at a laboratory to evaluate its performance. In a blind proficiency study, the laboratory personnel do not know that they are being tested.

prosecutor's fallacy. See transposition fallacy.

protein. Biologically important molecules made up of a linear string of building blocks called amino acids. The order in which these components are arranged is encoded in the DNA sequence of the gene that expresses the protein. See coding DNA.

pseudogenes. Genes that have been so disabled by mutations that they can no longer produce proteins. Some pseudogenes can still produce noncoding RNA.

quality assurance. A program conducted by a laboratory to ensure accuracy and reliability.

quality audit. A systematic and independent examination and evaluation of a laboratory's operations.

quality control. Activities used to monitor the ability of DNA typing to meet specified criteria.

random match. A match between a specific DNA profile and a second DNA profile drawn at random from the population. See also random-match probability.

random-match probability. The chance of a random match. As it is usually used in court, the random-match probability refers to the probability of a true match when the DNA being compared to the evidence DNA comes from a person drawn at random from the population. This random-true-match probability reveals the probability of a true match when the samples of DNA come from different, unrelated people.

random mating. The members of a population are said to mate randomly with respect to particular genes of DNA characteristics when the choice of mates is independent of the alleles.

rapid DNA. A fully automated process of developing a DNA profile from a sample without the need for a DNA laboratory or human intervention. A microfluidic lab-on-a-chip and associated software perform the extraction, amplification, separation, detection, and allele calling.

reads. Using next-generation "short-read" sequencing, DNA is broken into short fragments (typically 75–300 base pairs) that are amplified (copied) and then sequenced to produce "reads." Bioinformatic techniques then arrange the reads into a continuous genomic sequence. "Third-generation sequencing" methods permit "long-read" sequencing (> 10,000 bp).

recombination. In general, any process in a diploid or partially diploid cell that generates new gene or chromosomal combinations not found in that cell or in its progenitors.

reference population. The population to which the source of a trace-evidence sample is thought to belong.

relative fluorescent unit (RFU). See electropherogram.

replication. The synthesis of new DNA from existing DNA. See polymerase chain reaction.

restriction enzyme. Protein that cuts double-stranded DNA at specific base-pair sequences (different enzymes recognize different sequences). See restriction site.

restriction fragment length polymorphism (RFLP). Variation among people in the length of a segment of DNA cut at two restriction sites.

restriction fragment length polymorphism (RFLP) analysis. Analysis of individual variations in the lengths of DNA fragments produced by digesting sample DNA with a restriction enzyme.

restriction site. A sequence marking the location at which a restriction enzyme cuts DNA into fragments. See restriction enzyme.

reverse dot blot. A detection method used to identify SNPs in which DNA probes are affixed to a membrane, and amplified DNA is passed over the probes to see if it contains the complementary sequence.

ribonucleic acid (RNA). A single-stranded molecule "transcribed" from DNA. "Coding" RNA acts as a template for building proteins according to the sequences in the coding DNA from which it is transcribed. Other RNA transcripts can be a sensor for detecting signals that affect gene expression or can be a switch for turning genes off or on, or they may be functionless.

sequence-specific oligonucleotide (SSO) probe. Also, allele-specific oligonucleotide (ASO) probe. Oligonucleotide probes used in a PCR-associated detection technique to identify the presence or absence of certain base-pair sequences identifying different alleles. The probes are visualized by an array of dots rather than by the electropherograms associated with STR analysis.

sequencing. Determining the order of base pairs in a segment of DNA or RNA.

short tandem repeat (STR). A class of repetitive DNA resulting from multiple copies of virtually identical base-pair sequences, arranged in succession at a specific locus on a chromosome. The number of repeats varies from individual to individual. STRs are shorter than VNTRs.

single-locus probe. A probe that only marks a specific site (locus). A single-locus probe for a VNTR will produce one or two bands in an autoradiograph. Compare multilocus probe.

SNP (single nucleotide polymorphism). A substitution, insertion, or deletion of a single base pair at a given point in the genome. See polymorphism. Compare indel.

SNP chip. See chip.

Southern blotting. Named for its inventor, a technique by which processed DNA fragments, separated by gel electrophoresis, are transferred onto a nylon membrane in preparation for the application of biological probes.

thymine (T). One of the four bases, or nucleotides, that make up the DNA double helix. Thymine binds only to adenine. See nucleotide.

transcription. The process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). The mRNA transcript serves as a blueprint for protein synthesis during the process of translation. See messenger RNA.

transposition fallacy. Also called the prosecutor's fallacy, the transposition fallacy confuses the conditional probability of A given B with that of B given A.

Few people think that the probability that a person speaks Spanish given that he or she is a citizen of Chile equals the probability that a person is a citizen of Chile given that he or she speaks Spanish. Yet many court opinions, newspaper articles, and even some expert witnesses speak of the probability of a matching DNA genotype given that someone other than the defendant is the source of the crime-scene DNA as if it were the probability of someone else being the source given the matching profile. Transposing conditional probabilities correctly requires Bayes' rule.

true match. Two samples of DNA that have the same profile should match when tested. If there is no error in the labeling, handling, and analysis of the samples and in the reporting of the results, a match is a true match. A true match establishes that the two samples of DNA have the same profile. Unless the profile is unique, however, a true match does not conclusively prove that the two samples came from the same source. Some people use "true match" more narrowly, to mean only those matches among samples from the same source. Compare false match. See also match, random match.

variable number tandem repeat (VNTR). A class of RFLPs resulting from multiple copies of virtually identical base-pair sequences, arranged in succession at a specific locus on a chromosome. The number of repeats varies from individual to individual, thus providing a basis for individual recognition. VNTRs are longer than STRs.

whole genome sequencing. Determining the order of all the nucleotides in an individual organism's DNA. Usually accomplished with a massively parallel sequencing system.

window. See match window.

X chromosome. See chromosome.

Y chromosome. See chromosome.

References on DNA

Genetics and Genomics

John M. Archibald, Genomics: A Very Short Introduction (2018).
Jocelyn E. Krebs et al., Lewin's Genes XII (2018).
James D. Watson et al., Molecular Biology of the Gene (7th ed. 2014).

Forensic Genetics and Genomics

Jo-Anne Bright & Michael D. Coble, Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios (2020).
Forensic DNA Interpretation (John Buckleton et al. eds., 2d ed. 2016).
John M. Butler, Fundamentals of Forensic DNA Typing (2009); Advanced Topics in Forensic DNA Typing: Methodology (2011); Advanced Topics in Forensic DNA Typing: Interpretation (2015).
Silent Witness: Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters (Henry Ehrlich et al. eds., 2020).
Ian W. Evett & Bruce S. Weir, Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists (1998).
William Goodwin et al., An Introduction to Forensic Genetics (2d ed. 2011).
David H. Kaye, The Double Helix and the Law of Evidence (2010).
Erin E. Murphy, Inside the Cell: The Dark Side of Forensic DNA (2015).
National Research Council Committee on DNA Forensic Science: An Update, The Evaluation of Forensic DNA Evidence (1996).
National Research Council Committee on DNA Technology in Forensic Science, DNA Technology in Forensic Science (1992).
Royal Soc'y, Forensic DNA Analysis: A Primer for Courts (2017).

Reference Guide on Eyewitness Identification

THOMAS D. ALBRIGHT AND BRANDON L. GARRETT

Thomas D. Albright, Ph.D., is Professor and Director, Center for the Neurobiology of Vision, Conrad T. Prebys Chair in Vision Research at The Salk Institute for Biological Studies.

Brandon L. Garrett, J.D., is L. Neil Williams Professor of Law and Faculty Director, Wilson Center for Science and Justice at Duke University School of Law.

CONTENTS

FIGURES

Introduction

The Promise and Peril of Eyewitness Testimony

Witnessing the world and reporting what we have seen are among the most elemental of human abilities. We use these abilities to learn, to understand, and to assess what is true. We find things that we misplaced. We recognize our friends. Perhaps most importantly, we share our experiences with others. We testify to the beauty of the full moon rising, to the birth of our children, and to the terror of war.

Societies have long mined the power of this native human ability by recruiting eyewitnesses to help resolve disagreements about past events. In many cases, the outcome is of little consequence (did he swing the bat, or foul his opponent?), and we readily defer judgment to the witness. In other cases, we seek out eyewitness testimony to settle fraught conflicts over civil or criminal responsibility. A firsthand report of what happened and who was there can be extremely valuable, particularly when no physical evidence exists. But in court, a factfinder's decision rests not simply on what an eyewitness reports, but also on inference about the probability that the testimony is correct. It may not matter much if you misperceive, misremember, and misreport who was at the birthday party, but in a courtroom, being wrong about who assaulted you and hijacked the car can render a uniquely tragic form of injustice.

Eyewitness reports of criminal acts rarely involve deceit;[1] witnesses are generally well-intentioned people who offer a valued piece of information that they alone possess and believe to be true. Inaccurate testimony, should it occur, typically reflects an unwitting failure of human vision and memory, a failure that gets smoothed over, in the witness's telling, by candor and confidence. The result imperils the innocent,[2] leaves society at continued risk,[3] and undermines public trust. However, scientific advances shed light on the accuracy of eyewitness testimony. These advances have increasingly informed legal actors, including law enforcement, lawmakers, and courts. In this guide, we summarize several decades of scientific research and legal approaches, including those that incorporate scientific insights. We also discuss more recent discoveries and the challenges and opportunities posed by new technology.

1. A noteworthy exception is "understandable" deceit, in which a witness intentionally fails to identify a suspect (e.g., a gang member) for fear of retribution.

2. Brandon L. Garrett, Convicting the Innocent: Where Criminal Prosecutions Go Wrong (2011), https://doi.org/10.4159/harvard.9780674060982.

3. Jee Park, *Eyewitness Identification and Innocence*, 64 Loy. L. Rev. 669, 670 (2018) (explaining that "[o]f the 158 cases [of exonerations] where the true perpetrators were identified by DNA, these actual perpetrators went on to commit 150 additional violent crimes").

Definition of Eyewitness Identification

Eyewitness identification is defined in operational legal terms as "a naming or description by which one who has seen an event testifies from memory about the person or persons involved."[4] In the more analytical idioms of science, it is a type of visual object recognition task used for forensic purposes. Object recognition, in turn, is a ubiquitous function of human vision and memory, in which an observer must decide whether a currently viewed stimulus is the same as a previously viewed stimulus. People commonly perform such tasks in a variety of contexts and do so very successfully when the objects are familiar, such as one's coffee mug or the office stapler. By contrast, the majority of eyewitness reports in criminal cases are based on events in which a witness saw an unfamiliar person or persons during a single encounter, often briefly and under highly constrained conditions. Because people are poorer at identification of unfamiliar faces, relative to familiar ones,[5] the unfamiliar-face scenario presents a significant challenge to our criminal justice system. For purposes of this guide, eyewitness identification refers exclusively to the unfamiliar case.

A Word About Probabilities and Prediction

Probabilistic reasoning is a ubiquitous feature of courtroom decisions. Evidence is rarely certain; it varies continuously in the strength with which it is probative. In a criminal case, in order to convict, for example, a jury must infer whether eyewitness evidence is more likely under one hypothesis (the chef pulled the trigger) versus a different hypothesis (someone else pulled the trigger).

Analogously, a trial judge may apply probabilistic reasoning to make decisions about the accuracy and admissibility of evidence. The best approach to this problem is to assess conditions associated with the viewing and identification events that are known from prior empirical studies to increase or decrease accuracy. The result is a form of prediction in which the evidence is assigned (often implicitly) a probability of being true. From a continuous range of such probabilities, a judge would then apply a threshold, or decision criterion, to make a binary decision about whether to permit the evidence at trial.

Building on this framework of probabilistic inference, scientific approaches to the eyewitness accuracy problem fall into two broad categories: prospective and retrospective. The prospective approach uses novel techniques to *elicit* greater accuracy in proffered testimony. We review some of these techniques herein, but the problem faced by the courts is necessarily retrospective; the goal here is to

4. Black's Law Dictionary 894 (11th ed. 2019).

5. Bruce Vicki et al., *Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images*, 7 J. Exp. Psych.: Applied 207 (2001), https://doi.org/10.1037//1076–898x.7.3.207.

better *predict* the level of accuracy. This is a problem that the sciences of visual perception and memory are well equipped to address.

We stress these points at the outset because, as we will show, the assessment of the reliability of eyewitness evidence is not only the task of the jurors as fact-finders, but it is also central to the judge's task under the Supreme Court's 1977 *Manson v. Brathwaite* ruling on eyewitness evidence. Modern scientific research provides an empirical basis for assessing such probabilities and making predictions about the accuracy of eyewitness identifications, which is the subject of this guide.

Organization of This Reference Guide

The topic of eyewitness identification has been an intense focus of scientific research and legal policy development for decades. In this guide we have synthesized a vast amount of information from a variety of sources. We offer here a brief precis to assist with navigation of the text.

We begin with a generic description of the eyewitness and the task they are confronted with, which serves as an introduction to the problem faced by the criminal justice system. Casting the eyewitness as a biological information-processing "instrument" helps to focus on general factors that limit instrument performance under the constraints of the task.

With this orientation to the problem, we turn to accuracy prediction, which lies at the heart of any judicial decision about use of eyewitness testimony. We summarize the Supreme Court ruling in *Manson v. Brathwaite*, which recognized the significance of accuracy prediction and promoted a rule-based method for doing so. We follow this with a brief account of modern scientific approaches to prediction, which have ready application to the eyewitness problem and may improve upon the existing legal approach.

The next section provides a brief taxonomy of scientific questions that have been raised in research aimed at understanding eyewitness performance and use of eyewitness testimony by the courts. This taxonomy is followed by a dive into the science itself and the implications of this knowledge for assessment of the accuracy of eyewitness testimony.

The relevant science is rich and broad. We subdivide it into (1) basic science of vision and memory, with a focus on how these systems work and how their operational characteristics limit the accuracy of eyewitness reports; and (2) applied eyewitness science, which has experimentally manipulated variables commonly associated with the eyewitness experience, such as viewing conditions or identification procedures employed by the police, and assessed effects of these variables on identification performance. These variables—together with an understanding of how vision and memory work—provide a foundation for predicting the accuracy of eyewitness reports in criminal cases.

The final section of this guide turns to legal policy developments over the past half-century, which are intended to inform and regulate the use of eyewitness testimony. While much of this legal framework will be familiar to members of the judiciary, we show that these developments—which include court rulings, legislative actions, and changes in police practices—are increasingly rooted in the growing body of basic and applied science summarized herein.

The Eyewitness as an Information-Processing "Instrument"

From a scientific or technical perspective, it is useful to consider an eyewitness as a biological instrument that records, recovers, and reports previously encountered events—a sequence known generically as "information processing."[6] As for any functioning instrument, there are three pieces of knowledge needed to understand its utility for application: (1) How is it going to be used? (2) How does it work? and (3) *How well* does it work? In the following section, we briefly summarize the ways in which eyewitnesses are commonly used in criminal investigation and prosecution, and we provide an example from an actual criminal case to illustrate some of the strategies and pitfalls. In later sections on the basic sciences of vision and memory, and the applied science of eyewitness identification, we describe how the instrument works and how well it can perform the task at hand. This knowledge is critical for understanding the predictive relationship between variables associated with an eyewitness event and the accuracy of an identification.

How Is the Instrument Used?: A Taxonomy of Identification Procedures

The procedures used by law enforcement for eyewitness identification have become standardized through adherence to disseminated guidelines.[7] There are three basic procedures in use today: (1) showups, (2) photo arrays, and (3) live lineups. In addition, there are a variety of nonstandard identification procedures that are employed occasionally by police, or by eyewitnesses themselves.

6. "Information" here refers to content acquired through the senses, retrieved from memory, or derived from comparisons thereof, which improves the observer's ability to make predictions about the state of the world.

7. *See, e.g.*, U.S. Dep't of Just., Eyewitness Identification: Procedures for Conducting Photo Arrays (2017), https://perma.cc/TP4F-NXCC.

Showups

In a showup, which usually occurs at or near the crime location, police officers present a single live suspect to a witness. Showup procedures are only useful shortly after the crime, as when a person matching the witness's description is discovered in proximity to the crime scene and the witness's memory of the events is believed to be fresh. Any benefit gained by this proximity is countered, however, by the inherent suggestiveness of a showup: The procedure presents a witness with a single choice, which may cause the witness to infer that the police believe that the suspect is the perpetrator.[8] Because of this suggestiveness, showups have been, as the U.S. Supreme Court has put it, "widely condemned."[9]

Photo Arrays

Photo arrays are the most commonly used eyewitness identification procedure. In this case, the eyewitness is presented with six facial photographs,[10] each in a standard portrait view. One photo is that of a suspect; the other five photos are of "fillers"—people known to be innocent who are drawn from a preexisting database. The fillers serve to challenge recognition memory and reduce suggestiveness. Ideally, the filler faces should be chosen to match the witness's description of the perpetrator (a common police practice),[11] not the appearance of the suspect, because the witness's description is a direct reflection of what they have stored in memory from the crime scene.

Photo arrays are of two types: simultaneous and sequential. In a simultaneous presentation, officers display all facial photos at the same time, commonly in a two by three arrangement known as a "six-pack" photo array. In a sequential presentation, by contrast, photos are displayed one at a time.[12] Problems with suggestiveness in photo arrays may still arise if fillers are not carefully chosen. For example, a lineup that includes some fillers who do not match the witness's description of the perpetrator reduces the number of sensible choices.

8. *See* Nancy Steblay et al., *Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta–Analytic Comparison*, 27 Law & Hum. Behav. 523, 523–24 (2003), https://doi.org/10.1023/a:1025438223608.

9. Stovall v. Denno, 388 U.S. 293, 302 (1967).

10. Six is the standard used in the United States today, where it is deemed sufficient to challenge recognition memory. But there is not universal agreement on this number: In the United Kingdom, for example, lineups ("identity parades") are commonly composed of nine faces.

11. Police Exec. Research Forum, A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies (2013), https://perma.cc/REX2-Y29T.

12. Nat'l Research Council, Identifying the Culprit: Assessing Eyewitness Identification 24 (2014) [hereinafter Identifying the Culprit], https://perma.cc/4246-37D2.

Live Lineups

Before the ready availability of standardized facial photographs, live lineups were used. In this procedure, the suspect and fillers are presented in person, either as a group (simultaneous) or one at a time (sequential). While some agencies still use live lineups, they are less common today because police may need probable cause to place a nonconsenting person in a lineup, presence of counsel may be required, and it can be practically difficult to find live fillers who fairly resemble the witness's description of the perpetrator.[13]

Nonstandard Identification Procedures

A range of additional procedures may be used in situations in which officers do not have a suspect, including showing mug books or sets of photographs or asking the eyewitness to help prepare a composite image or drawing of a culprit. Police will, on occasion, display a single photograph to a witness in an effort to confirm the identity of a person of interest. Police typically limit this "confirmatory photograph" method to situations in which the person is previously known to or acquainted with the witness. Police may sometimes take a witness to view people in the field, perhaps at a location the culprit is known to frequent. Finally, eyewitnesses may try to make identifications not arranged by police, in person or on social media, offender registries, or other online image collections. All of these situations are less controlled than a lineup procedure and potentially suggestive.

An Illustrative Case

To help illustrate the process by which eyewitness evidence emerges and its potential unfolds, and to illuminate the manifold variables that come into play when a trier of fact is evaluating the accuracy of the evidence, we begin with an example from the real world. We reference this example occasionally in the remainder of the guide in order to place scientific discussions in context.

As a juror in a lengthy criminal trial, you are listening to the prosecution recount how while walking to a friend's house, a 16-year-old student was abruptly abducted, dragged into roadside bushes, and sexually assaulted by an unknown man.[14] After repeatedly punching him in the face, the victim managed to break free and run away. A passerby, in his car, saw the face of the attacker. Later, during

13. *Id.* at 25.

14. Thomas D. Albright, *Why Eyewitnesses Fail*, 114 PNAS 7758, 7758–64 (2017) [hereinafter Albright, *Why Eyewitnesses Fail*], https://doi.org/10.1073/pnas.1706891114.

the trial, you listen to the victim's and passerby's testimony. When asked if the victim was confident that the defendant sitting in the courtroom was the culprit, the victim replies: "Yes, I'm sure . . . I will never forget what he looks like." The passerby confirms the identification, and confirms that he has no doubts at all. As a juror, your instinct is to believe the two witnesses who observed the crime and confidently conveyed their testimony. How could you not? One saw—and courageously punched—the assailant, and the other saw the assailant just moments afterward. How could either be wrong about the attacker's identity?

The situation just described is the case of Uriah Courtney, who was sentenced to life in prison. After Courtney served eight years, however, the court granted a motion to retest DNA from the victim's clothing. The results excluded Courtney, and he was released in 2013. The DNA evidence was linked, through a hit in a DNA database, to a man who lived near the crime scene. The wrongful conviction of Uriah Courtney naturally raises questions about the accuracy of the victim's testimony and points to three general factors that often undermine criminal investigations and prosecutions based on eyewitness testimony: uncertainty, bias, and overconfidence. In simple terms, appreciation of the extent to which these three factors are in play can help improve inferences about the accuracy of eyewitness testimony.

One cause of uncertainty is the fact that the second witness and Courtney were of different races and ethnicities. It is well known that people generally have greater difficulty discriminating faces of a race different from their own (relative to same-race face discrimination),[15] which manifests as disproportionately large eyewitness errors in cross-racial identifications and has significant implications for racial justice.

There were also potential sources of bias built into the identification procedure used in Uriah Courtney's case. For example, the police showed a traditional simultaneous photo array to the two eyewitnesses, who had described the culprit as white, in his mid-twenties, with brown hair and a goatee. Only two of the five fillers in this lineup had any facial hair at all—and Courtney's goatee was the most conspicuous. This lineup configuration, in conjunction with the witnesses' descriptions, naturally affected the statistics of the identification process, leading to a high prior likelihood—a bias—for picking Courtney, which both eyewitnesses did. Additional potential for bias was introduced by the fact that the officer administering the lineups knew which participant was the suspect, which allowed for inadvertent behavioral signals that could have influenced the witness's decisions.

15. Rankin W. McGugin et al., *Race-Specific Perceptual Discrimination Improvement Following Short Individuation Training with Faces*, 35 Cognitive Sci. 330–47 (2011), https://doi.org/10.1111/j.1551-6709.2010.01148.x; Hoo Heat Wong, Ian D. Stephen & David R. T. Kebble, *The Own-Race Bias for Face Recognition in a Multiracial Society*, 11 Frontiers Psych. 208 (2020), https://doi.org/10.3389/fpsyg.2020.00208.

Finally, during the Courtney trial the witnesses expressed marked confidence in their identifications, which likely moved the jury to convict, despite the fact that upon the victim's first viewing of a photo lineup, she was tentative, telling police that the defendant's photo was "most similar," but that she was "not sure," and had a confidence level of 60%.[16] By the time of the trial, her confidence had become inflated, while the accuracy of her testimony remained unchanged.

Uriah Courtney's wrongful conviction is but one in a lengthy and growing list that involved eyewitness misidentifications, in which accuracy of eyewitness testimony was incorrectly inferred by the court. Of the over 375 individuals in the United States who have been exonerated as of this writing based on postconviction DNA testing, approximately 70% were convicted, in part, based on eyewitness misidentifications.[17] We do not know how often eyewitness identifications are conducted, but according to one estimate, they may be used in tens of thousands of cases a year.[18]

Predicting the Accuracy of Eyewitness Identification

As emphasized above, it is impossible for a user—law enforcement or the courts—to rationally and responsibly employ eyewitness testimony without information about the probability that the testimony is correct. But where does that information come from? There is no oracular source, and witness statements about their own accuracy are naturally suspect. As with other indeterminate systems, such as medical prognoses or weather forecasting, we must draw from other pieces of information to predict eyewitness accuracy.

In this section we briefly review principles of accuracy prediction developed by law and science communities. Before doing so, we highlight a distinction between (1) estimates of the probability that witness testimony would be accurate under a particular set of known conditions, and (2) the probability that a specific witness is accurate.[19] As we will show, knowledge gained from both

16. *See* Uriah Courtney, Innocence Center, https://theinnocencecenter.org/case/uriah-courtney/.

17. *See* Brandon L. Garrett, Convicting the Innocent: DNA Exoneration Resource Database, *available* at https://www.convictingtheinnocent.com (providing detailed information concerning DNA exoneration cases); Garrett, *supra* note 2 (summarizing DNA exoneration data); Brandon L. Garrett, *Judging Innocence*, 108 Colum. L. Rev. 55 (2008) ("The vast majority of the exonerees (79%) were convicted based on eyewitness testimony.").

18. Identifying the Culprit, *supra* note 12 (citing Alvin G. Goldstein, June E. Chance & Gregory R. Schneller, *Frequency of Eyewitness Identification in Criminal Cases: A Survey of Prosecutors*, 27 Bull. Psychonomic Soc'y 71, 73 (1989) https://doi.org/10.3758/bf03329902); Garrett, *supra* note 2, at 50.

19. *See* David L. Faigman, John Monahan & Christopher Slobogin, *Group to Individual (G2i), Inference in Scientific Expert Testimony*, 81 U. Chi. L. Rev. 417, 417–80 (2014).

basic and applied sciences makes it possible, in principle, to offer quantitative estimates of accuracy under known conditions. By contrast, we would discourage attempts by experts to apply a numerical probability to a *specific* witness, recognizing that doing so wades into the province of the court, and there are numerous unknown variables that bear on individual cases.

Manson v. Brathwaite: Accuracy Prediction by the Court

In its landmark 1977 *Manson v. Brathwaite* ruling on the use of eyewitness evidence, the U.S. Supreme Court explicitly promoted a predictive strategy for assessing the accuracy of eyewitness reports, which was intended to provide orderly grounds for decisions about the admissibility of the evidence under the Due Process Clause. This strategy involves application of a five-factor test, in which specific conditions of witnessing and reporting are compared to five conditional standards—often termed "reliability criteria"—believed to be correlated with (and potentially causal of) the probability that a witness's testimony is correct. We briefly review these criteria and the logic behind them here because they provide the constitutional judicial standard for admissibility of eyewitness evidence (although as discussed, federal and state rules of evidence also govern eyewitness testimony in court, as well as state statutes). For each, we suggest how growing scientific knowledge of human information processing (reviewed in greater detail below) supports a predictive strategy and may bring us closer to precisely estimating the probability that eyewitness testimony is correct.[20]

The first *Manson* reliability criterion—"the opportunity of the witness to view the criminal at the time"—is intended to address the question of visual fidelity, or accuracy of a witness's perceptual experience. An "opportunity to view" (commonly termed "line-of-sight visibility" by the scientific community[21])

20. The disconnect between the *Manson* "reliability criteria" and modern sciences of vision and memory has been noted by others: Gary L. Wells & Deah S. Quinlivan, *Suggestive Eyewitness Identification Procedures and the Supreme Court's Reliability Test in Light of Eyewitness Science: 30 Years Later*, 33 L. & Hum. Behav. 1, 1–24 (2009); Randolph N. Jonakait, *Reliable Identification: Could the Supreme Court Tell in* Manson v. Brathwaite*?*, 52 U. Colo. L. Rev. 511 (1981); Ruth Yacona, Manson v. Brathwaite: *The Supreme Court's Misunderstanding of Eyewitness Identification*, 39 J. Marshall L. Rev. 539 (2006); Laura Smalarz et al., *Psychological Science on Eyewitness Identification and the U.S. Supreme Court: Reconsiderations in Light of DNA-Exonerations and the Science of Eyewitness Identification*, *in* The Witness Stand and Lawrence S. Wrightsman, Jr. (Cynthia Willis-Esqueda & Brian Bornstein eds., 2016).

21. M.L. Benedikt, *To Take Hold of Space: Isovists and Isovist Fields*, 6 Environ. & Planning B: Urb. Analytics & City Sci. 1 (1979), https://doi.org/10.1068/b060047.

is essential, to be sure, but the current state of vision science, much of which is focused on the very question of visual sensitivity and fidelity, is such that estimates of accuracy can now be made beyond the limiting case of what is within an observer's line of sight. In particular, observers do not simply view or not view things in their environment; there is a predictive relationship between variables that introduce uncertainty in visual experience and the accuracy of observer reports.

The second *Manson* criterion—"the witness' degree of attention"—acknowledges that attention promotes greater visual fidelity and certainty of measurement, and is thus predictive of accuracy. Modern science informs our understanding of this variable as well. Specifically, attention is a limited resource; direction of attention to one part of a visual scene routinely precludes attention to other parts. Thus, the "degree of attention" is primarily meaningful in the context of what is being attended to. As detailed under discussions of scientific research (see sections titled "Visual attention" and "Weapons focus" below), a high degree of attention to a weapon, for example, will naturally draw attentional focus away from the face, introducing uncertainty in visual experience of the face and reduction of identification accuracy. Building upon this information, it is now possible to make refined predictions of accuracy based on both the degree and the likely target of attention.

The third *Manson* criterion—"the accuracy of [the witness's] prior description of the criminal"—is based on the premise that a witness's ability to recall what the perpetrator looked like ("accuracy of his prior description") can be used to infer the probability that an identification made by the witness is correct. This is a valid premise for familiar objects, on which attention has been focused, perceptual processing is commonly "deep," and the level of detail recalled is significant.[22] By contrast, modern science reveals that the correlation between recall and recognition is not strong under conditions in which unfamiliar faces are witnessed under viewing conditions constrained by time, distance, lighting, stress, and poorly focused attention.[23] All of which suggests that the predictive value of a good description may be substantially limited by the conditions of viewing.

The fourth *Manson* criterion—"the level of certainty demonstrated at the confrontation"—is rooted in the simple premise that "level of certainty," or confidence, demonstrated at the "confrontation" (at the time of the initial identification test) is a predictor of eyewitness identification accuracy. This premise matches human intuition but, as we discuss in some detail below, modern science has shown that the relationship between confidence and accuracy of recognition

22. Fergus I.M. Craik & Robert S. Lockhart, *Levels of Processing: A Framework for Memory Research*, 11(6) J. of Verbal Learning & Verbal Behav. 671–84 (1972).

23. Melissa Pigott & John C. Brigham, *Relationship Between Accuracy of Prior Description and Facial Recognition*, 70 J. Applied Psych. 547 (1985), https://doi.org/10.1037//0021-9010.70.3.547.

memory is nuanced. Under some well-defined conditions,[24] the initial confidence report may correlate with the probability that the testimony is accurate.[25] As we see from the Courtney case, the predictive value of this measure can grow worse with the passage of time, largely owing to outside influences (e.g., reinforcement from media and counsel) that boost confidence but naturally have no effect on accuracy, reaching a nadir at the time of the in-court identification.

The fifth *Manson* criterion—"the length of time between the crime and the confrontation"—acknowledges that time-dependent memory loss increases uncertainty of recalled information, which means that memory retention time is inversely correlated with accuracy. Because accuracy is a continuous function of time and admissibility decisions are binary, a judge must apply a time threshold for evaluation. But where does that threshold come from? Scientific studies of the time–accuracy relationship show that memory loss can be modeled as a power law function, by means of which "not only can the percentage of remaining memory strength be determined for any retention interval, but this strength estimate can be translated into an estimated probability of being correct on a fair lineup of a specified size."[26] In other words, modern science can provide a quantitative, empirical foundation for the time threshold that a judge employs when evaluating evidence by the fifth *Manson* criterion.

The day-to-day judicial approach to eyewitness evidence has been largely driven over the past half-century by the predictive framework of *Manson v. Brathwaite*, at least in the federal courts. State courts, as we will discuss, have introduced more recent changes, through legislation and judicial rulings, seeking to more directly incorporate scientific insights into approaches toward eyewitness evidence. In federal courts, as in state courts, scientific insights have influenced the interpretation of the *Manson* factors and other related approaches, such as the use of expert witnesses to explain eyewitness evidence to the jury, and the use of jury instructions regarding eyewitness evidence. In the meantime, the scientific community has continued to weigh in on the important matter of inferring the probability that eyewitness testimony is correct, based primarily on new scientific discoveries in the area of human information processing and the use of new statistical tools.

24. "Pristine" lineup conditions include choosing lineup fillers such that they similarly match the witness's description of the perpetrator, making the lineup "fair," and imposing limits on biasing factors such as nonblinded lineup administration.

25. John T. Wixted & Gary L. Wells, *The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis*, 18 Psych. Sci. Pub. Int. 10 (2017), https://doi.org/10.1177/1529100616686966.

26. Kenneth A. Deffenbacher et al., *Forgetting the Once-Seen Face: Estimating the Strength of an Eyewitness's Memory Representation*, 14 J. Experimental Psych.: Applied 139, 142 (2008), https://doi.org/10.1037/1076-898x.14.2.139.

Accuracy Prediction by the Scientific Community

Motivated by societal concerns over wrongful convictions, which began to mount significantly with the development of efficient and effective procedures for forensic genotyping (starting approximately a decade after the Supreme Court's 1977 *Manson* ruling), a number of science–law collaborations emerged to explore how new scientific knowledge and tools might improve collection and use of eyewitness testimony.[27] The scientific foundations continue to evolve and sharpen, as do various recommendations for practice and reform.

In 2013, the National Research Council (NRC) convened a panel of experts to consider the use and validity of eyewitness identification for forensic purposes. The panel was composed of individuals representing a variety of relevant fields, including the scientific study of visual perception and memory, sociology, statistics, law, and policing. After reviewing evidence from many sources, the panel released a comprehensive report in the fall of 2014.[28]

Other organizations have subsequently performed reviews. Particularly noteworthy and commendable is the 2020 scientific review paper from the American Psychology-Law Society,[29] which covers much of the same scientific ground as the NRC report, but concentrates almost exclusively on recommendations to law enforcement for planning, designing, and conducting eyewitness identification procedures. In the remainder of this section, we focus primarily on the NRC report and its recommendations for law enforcement, for the courts, and for the advancement of eyewitness science.

The NRC report recognized the great value that eyewitness testimony can bring to criminal proceedings and the concerted efforts of law enforcement and

27. Gary L. Wells et al., *Eyewitness Identification Procedures: Recommendations for Lineups and Photospreads*, 22 Law & Hum. Behav. 603 (1998), https://doi.org/10.1023/a:1025750605807; U.S. Dep't of Just., Off. of Just. Programs & Nat'l Inst. of Just., Eyewitness Evidence: A Guide for Law Enforcement (1999), https://perma.cc/Y9EA-B4W3; John W. Turtle, R.C. Lindsay & Gary L. Wells, *Best Practice Recommendations for Eyewitness Evidence Procedures: New Ideas for the Oldest Way to Solve a Case*, 1(1) Canadian J. of Police & Sec. Servs. 5 (2003); *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, 92 Temp. L. Rev. 1, 6 (2019), https://perma.cc/Q4CH-5NLU; Gary L. Wells et al., *Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence*, 44 Law & Hum. Behav. 3 (2020), https://doi.org/10.1037/lhb0000359; Margaret Bull Kovera et al., *Science-Based Recommendations for the Collection of Eyewitness Identification Evidence*, 58 Ct. Rev. 130 (2022).

28. Identifying the Culprit, *supra* note 12 (one author of the present guide served as co-chair of the National Resarch Council consensus report committee, and the other author served as a committee member).

29. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

legal professionals. At the same time, however, the NRC report identified factors that may lead to wrongful conviction and made recommendations for improvement in three areas: law enforcement procedure, use of eyewitness evidence in the courts, and scientific research. With respect to eyewitness performance, the NRC recommendations are both prospective, with an eye toward eliciting greater eyewitness accuracy through procedural reforms, and retrospective, with a focus on identifying and quantifying variables that offer predictive insight into the probability that eyewitness testimony is correct.

These recommendations have been embraced by both scientific and legal communities and have had, in a few short years, a significant positive impact on the field. The NRC report itself, which is freely available from the National Academies Press, has been downloaded more than 17,000 times and has elicited rich discussion, new scientific research, and improved legal practice. In 2017, U.S. Deputy Attorney General Sally Yates issued new Department of Justice (DOJ) guidelines for the conduct of eyewitness investigations,[30] which hew tightly to the recommendations and language of the NRC report. Though the charter of the independent NRC defines and limits its work to be strictly advisory—unlike the DOJ, the NRC has no enforcement powers—we briefly highlight the NRC recommendations on eyewitness identification in this section, as they offer a scholarly complement to the predictive framework established by the Supreme Court in 1977.

Law Enforcement

The five NRC recommendations for law enforcement are intended to foster "adherence to guidelines that are consistent with scientific standards for data and reporting."[31] These include:

1. Train all law-enforcement officers in eyewitness identification
2. Implement double-blind lineup and photo array procedures
3. Develop and use standardized witness instructions
4. Document witness confidence judgments
5. Videotape the witness identification process

These law enforcement recommendations are aimed prospectively at improving the quality of evidence elicited from eyewitnesses and presented to

30. *See* U.S. Dep't of Just., *supra* note 7.
31. *See* Identifying the Culprit, *supra* note 12, at 105.

the court. Adherence to these recommendations can reduce the likelihood that uncertainty, bias, and overconfidence come into play, and make it easier to retrospectively predict the probability that an identification hypothesis is true.

The Courts

The four NRC recommendations for the courts encourage adoption of procedures intended to help courts distinguish evidence that does or does not have high predictive validity. These include:

1. Conduct pretrial judicial inquiry
2. Make juries aware of prior identifications
3. Use scientific framework expert testimony
4. Use jury instructions as an alternative means to convey information

Pretrial inquiry offers an opportunity for the judge to indulge in a careful consideration of variables (those identified by *Manson*, or others) that might serve retrospectively as good predictors of the accuracy of an identification. Use of scientific framework testimony (testimony about relevant scientific facts and principles, without explicit application to the case in question) and use of "clear and concise" jury instructions both provide similar opportunities for jurors to consider variables that are good predictors of the accuracy of an identification.[32] Finally, the recommendation that juries be made aware of prior identifications is intended to establish the shortest retention interval, reveal inconsistencies, and combat the influence of confidence inflation on jury deliberations. All of which should improve the quality of evidence used retrospectively to predict the accuracy of an identification.

Scientific Research

Finally, the NRC report made several recommendations for additional scientific research, recommendations aimed at developing strategies to both prospectively elicit more accurate testimony from eyewitnesses, and—for retrospective purposes—to gain greater understanding of the causal relationships between variables associated with witnessed events and accuracy of the resulting

32. Jury instruction was an explicit recommendation of the NRC eyewitness committee. Nonetheless, as summarized below in the context of scientific studies conducted in the laboratory, some research suggests that jury instructions are not always effective at preventing wrongful convictions.

testimony. As summarized in detail in the next sections, this scientific research has progressed at a rapid pace, and particularly so since the publication of the NRC report.

Scientific Questions Regarding Eyewitness Identification

The past half-century has seen the accumulation of a vast trove of scientific knowledge bearing on the validity of eyewitness testimony. This knowledge has greatly enhanced understanding of the capabilities and limitations of this form of evidence, and has advised new policies and practices. In the foregoing sections we distilled some of this science to provide ready insights for common judicial decisions. In the following sections we take a deep dive into the scientific discoveries that informed our summary, with the conviction that this knowledge will serve as a comprehensive reference that allows for greater judicial flexibility and understanding of when, why, and with what probability eyewitnesses provide accurate testimony. First, we briefly outline the types of scientific questions that are commonly addressed. This is followed by a summary of findings from: (1) basic laboratory studies that ask how brain systems for vision and memory work; (2) applied laboratory studies that specifically focus on human performance in an eyewitness task; (3) field studies that test the effectiveness of new eyewitness tasks in real criminal justice settings; and (4) studies that explore the efficacy of communication between the eyewitness and the trier of fact.

Most scientific studies of eyewitness identification conducted over the past half-century have focused on three central questions:

1. Can we improve prediction of eyewitness identification accuracy? When an eyewitness makes an identification, the question naturally of foremost importance for the criminal justice system is: What is the probability that the eyewitness is correct? This is the retrospective question, which focuses on variables associated with things past (e.g., conditions of witnessing and collection of evidence by law enforcement) that can be used to infer the accuracy of eyewitness reports. Nearly all of the science relevant to eyewitness evidence has addressed this question, either directly or indirectly.

2. Are there procedural improvements that can increase accuracy? The second, prospective, question seeks strategies that the criminal justice system might use to increase the probability of accurate testimony. The scientific literature is thick with empirical comparisons of performance (e.g., accuracy) elicited by different identification procedures, such as showups versus lineups, or

simultaneous versus sequential lineup presentations. The answer may favor the use by police of one procedure over the other. Most importantly for the present context, awareness by the trier of fact of these prospective improvements may serve retrospectively to help infer the probability that eyewitness testimony is correct.

3. Can we educate jurors to be better evaluators of accuracy?

The third scientific question addresses the outsized influence of eyewitness testimony on decisions by juries. The importance of this issue can be appreciated from a scientific perspective in which the use of eyewitness evidence by the courts is viewed as a process of communication: The eyewitness is the "transmitter" of information. The jury, conversely, is the "receiver" of information. If that receiver is flawed by an inability to interpret the signals that were transmitted—either by uncritical acceptance or by lack of technical knowledge—then the fidelity of information transmitted (the accuracy of the testimony) makes little difference to the outcome. It follows that the quality of judicial decisions might benefit from plain language explanations of (1) the underlying processes of vision and memory, (2) the ways that those processes can fail, contrary to intuition, and (3) why eyewitness candor and confidence are not necessarily predictive of accuracy. Carefully worded jury instructions and scientific expert testimony have been implemented in a number of jurisdictions. While some evidence challenges the effectiveness of this approach,[33] recent experiments suggest that instructions may be more effective if they explain the reasons *why* caution is critical when evaluating eyewitness accuracy, rather than simply directing caution.[34] We summarize the relevant science below.

33. Amanda N. Bergold et al., *Eyewitnesses in the Courtroom: A Jury-Level Experimental Examination of the Impact of the* Henderson *Instructions*, 17 J. Experimental Criminology 433, 433–55 (2021), https://doi.org/10.1007/s11292-020-09412-3; Marlee Kind Berman, *Eyewitness Identification Jury Instructions: Do They Enhance Evidence Evaluation?*, C.U.N.Y. Acad. Works (2015); Angela M. Jones & Amanda N. Bergold, *Examining the Effectiveness of the Henderson Eyewitness Instructions*, 13 J. Forensic Psych. Prac. 1, 1–24 (2017), https://doi.org/10.1007/s11292-016-9279-6; Brandon L. Garrett et al., *Factoring the Role of Eyewitness Evidence in the Courtroom*, 17 J. Empirical Legal Stud. 556, 556–79 (2020), https://doi.org/10.1111/jels.12259; Marlee Kind Dillon et al., *Henderson Instructions: Do They Enhance Evidence Evaluation?*, 17 J. Forensic Psych. Rsch. & Practice 1 (2017), https://doi .org /10 .1080 /15228932 .2017 .1235964; Angela M. Jones et al., *Comparing the Effectiveness of* Henderson *Instructions and Expert Testimony: Which Safeguard Improves Jurors' Evaluations of Eyewitness Evidence?*, 13 J. Exp. Crim. 29, 29–52 (2017), https://doi.org/10.1007/s11292-016-9279-6; Cindy E. Laub, Christopher D. Kimbrough & Brian H. Bornstein, *Mock Jurors' Perceptions of Eyewitnesses Versus Earwitnesses: Do Safeguards Help?*, 34(2) Am. J. Forensic Psych. 33 (2016); Athan P. Papailiou, David V. Yokum & Christopher T. Robertson, *The Novel New Jersey Eyewitness Instruction Induces Skepticism but Not Sensitivity*, 10(12) PLoS One e0142695 (2015), https://doi.org/10.1371/journal.pone .0142695.

34. Brandon L. Garrett et al., *Sensitizing Jurors to Eyewitness Confidence Using "Reason-Based" Judicial Instructions*, 12(1) J. Applied Rsch. Memory & Cognition 141 (2022), https://doi.org/10.1037 /mac0000035.

Basic Science of Vision, Memory, and Choice

Vision, memory, and decision-making have been subjects of rigorous scientific investigation since the middle of the nineteenth century, with particularly rapid advances over the past 50 years. In simple terms, the goal has been to understand how and how well the machine works. If we're considering the purchase of a new car, violin, or patio heater, the first thing we want to know is what are the use conditions under which it performs well or fails. Unlike engineered products, however, human vision and memory do not come with performance specs and a user manual. The purpose of basic research in this area is to obtain information that can enable prediction of human performance under specific conditions, such as those associated with witnessing and reporting a crime.[35] In the following sections, we summarize what has been learned from this basic science.

How Vision Works

Vision is usefully understood as the process of detecting informative signals about the external world and using those signals to recognize objects, make decisions, and guide behavior.[36] Our focus here is on aspects of visual function that place constraints on what can be sensed and perceived. In two sections, we address (1) characteristics of vision that place upper bounds on performance, and (2) vision as a constructive process that seeks to overcome input noise and uncertainty.

Upper Bounds on Visual Performance

In this section we review some specific aspects of visual processing that place upper bounds on the quantity and quality of visual information available under defined viewing conditions. This knowledge offers a starting point for evaluating the probability that an eyewitness could have seen what they said they saw.[37]

35. *See* Identifying the Culprit, *supra* note 12; Albright, *Why Eyewitnesses Fail*, *supra* note 14, at 7758–64; Marc Green et al., Forensic Vision with Application to Highway Safety (3rd ed. 2008).

36. Charles D. Gilbert, *The Constructive Nature of Visual Processing*, *in* Principles of Neural Science 496 (Eric R. Kandel et al. eds., 6th ed. 2021); Charles D. Gilbert, *Intermediate-Level Visual Processing and Visual Primitives*, *id.* at 545; Thomas D. Albright & Winrich A. Freiwald, *High-Level Visual Processing: From Vision to Cognition*, *id.* at 564; Thomas D. Albright, *On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception*, 74 Neuron 227 (2012), https://doi.org/10.1016/j.neuron.2012.04.001.

37. *See* Identifying the Culprit, *supra* note 12; Albright, *Why Eyewitnesses Fail*, *supra* note 14; Green et al., *supra* note 35.

Visual sensitivity

At the earliest stages of visual processing there are several factors that limit the visual information available for accurate perception and object recognition. Some of these factors are inherent to the visual system and largely uncontrollable, such as scattering of light by the fluid and tissues of the eye, and can be exacerbated by common observer-specific visual deficits, such as myopia, poor contrast sensitivity, or color blindness. Other factors are associated with viewing conditions, such as viewing duration and ambient illumination, which predictably influence the quantity of information that a viewer gains from a visual scene, and thus the fidelity of experience.

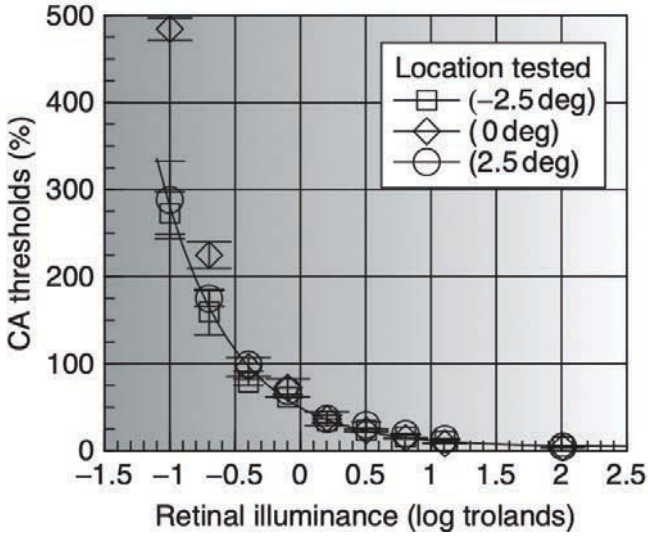
Anyone with a functioning visual system will report that their ability to resolve fine details of a viewed scene declines steeply in dim light; this is why we use flashlights, and why we cannot read a restaurant menu under candlelight. The basic science of vision puts numbers behind this experience, which can be used to predict the accuracy of eyewitness testimony.[38] Figure 1 shows how this effect can be quantified. The y-axis plots the amount of luminance contrast between an image feature and the background—for example, the brightness difference between a small black letter on a white page—that is required to resolve the image feature.[39] The x-axis plots ambient illumination. The plot shows that under very dim light (near total darkness; left side of x-axis), the image feature can be resolved only if it is extremely bright relative to background. By contrast, at high levels of illumination (daylight; right side of x-axis), feature resolution is easy. But there are intermediate levels of illumination (moonlight or candlelight; middle of x-axis), which are frequently encountered by human observers and make resolution very difficult.

The plot in Figure 1 represents one of the operating characteristics of human vision. In conjunction with measured values of ambient illumination present at a witnessed crime scene—knowable, in principle, from measurements of light sources, such as daylight, moonlight, street lights, illuminated windows, or signage—these operating characteristics can be used to predict the fidelity of experience and, in the spirit of *Manson*, assess with respectable precision whether a witness had an opportunity "to view the criminal at the time." It is precisely for this reason that human operating characteristics—hard-won products of basic research—are such a valuable resource for accuracy predictions in the eyewitness context.

38. Percy W. Cobb, *The Influence of Illumination of The Eye on Visual Acuity: I. Introductory and Historical*, 29 Am. J. Physiology 76 (1911), https://doi.org/10.1152/ajplegacy.1911.29.1.76; Simon Shlaer, *The Relation Between Visual Acuity and Illumination*, 21 J. Gen. Physiology 165 (1937), https://doi.org/10.1085/jgp.21.2.165.

39. J.L. Barbur & A. Stockman, *Photopic, Mesopic and Scotopic Vision and Changes in Visual Performance*, *in* Encyclopedia of the Eye, vol. 3, 323 (Darlene A. Dartt et al. eds., 2010), https://doi.org/10.1016/b978-0-12-374203-2.00233-5.

Figure 1. Sensitivity to luminance contrast as a function of overall illumination.



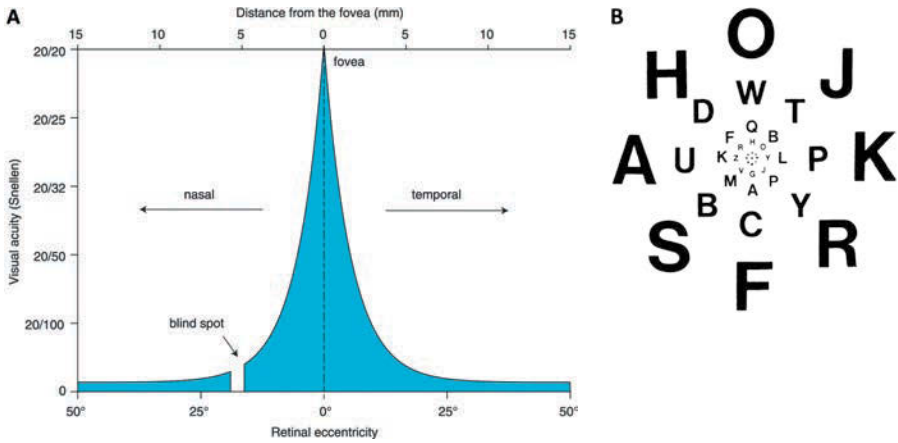
Source: Used with permission of Elsevier Science & Technology Journals, from J. L. Barbur & A. Stockman, *supra* note 39, at 323 (permission conveyed through Copyright Clearance Ctr., Inc.).

Another compelling example comes from quantitative information about visual acuity.[40] In Figure 2A, the y-axis is a measure of visual acuity. The x-axis is a measure of angular distance from the center of gaze. Normal maximum acuity is 20/20, which is achieved at the center of gaze (0 deg "retinal eccentricity" in this plot). At a viewing angle of 2 degrees from the center of gaze (about the width of a thumb at arm's length), acuity declines by half. The perceptual consequences of this decline in acuity are illustrated in Figure 2B, where letters are scaled to be equally resolvable when directing eye gaze to the center of the image. Under time-limited, distracting, and stressful viewing conditions associated with viewing a typical crime, off-axis gaze direction—looking directly at other compelling features (such as weapon, a bleeding victim, or a second perpetrator)—can be expected to place severe constraints on the ability to detect, resolve, and memorize the face of the primary culprit, because that face image will be registered only by low-acuity regions of the sensory field.[41] It follows that if one has facts about the scene, such as viewing position, distances, and angles, in conjunction

40. S.M. Anstis, *A Chart Demonstrating Variations in Acuity with Retinal Position*, 14(7) Vision Rsch., 589 (1974).

41. Haus Strasburger, Ingo Rentschler & Martin Jüttner, *Peripheral Vision and Pattern Recognition: A Review*, 11 J. Vision 13 (2011), https://doi.org/10.1167/jov.10.1167/11.5.13.

Figure 2. Visual acuity declines predictably with distance from center of gaze. *Panel A:* Plot of acuity as a function of eccentricity. *Panel B:* Perceptual consequences of acuity loss.



Source (Panel A): Stanley Lambertus et al., *Highly Sensitive Measurements of Disease Progression in Rare Disorders: Developing and Validating a Multimodal Model of Retinal Degeneration in Stargardt Disease*, 12 PLoS One (2017), https://doi.org/10.1371/journal.pone.0174020. Copyright: © 2017 Lambertus et al. This is an open access article distributed under the terms of the Creative Commons Attribution License. (Panel B): Used with permission of Elsevier Science & Technology Journals, from Anstis, *supra* note 40, at 589–92 (permission conveyed through Copyright Clearance Ctr., Inc.).

with the acuity function shown above in Figure 2A, it should be possible to draw probabilistic inferences about the degree to which the witness had an opportunity "to view the criminal," bearing (per *Manson*) on the likelihood that the witness testimony is accurate.[42]

The ability to detect, resolve, and interpret image content is further affected in known ways by a myriad of other environmental factors, such as atmospheric conditions (rain, fog), glare, and object surface reflectance.[43] "Crowding," in which objects in a visual scene are closely spaced, significantly reduces the ability

42. There is a large literature on this same topic applied to the ability of field athletes (e.g., soccer, football, tennis) to perceive critical events as a function of gaze angle, e.g., Karl Marius Aksum et al., *What Do Football Players Look At? An Eye-Tracking Analysis of the Visual Fixations of Players in 11 v. 11 Elite Football Match Play*, 11 Frontiers Psych. (2020), https://doi.org/10.3389/fpsyg.2020.562995; Francisco Belda Maruenda, *Can the Human Eye Detect an Offside Position During a Football Match?*, 329 BMJ 1470–72 (2004).

43. Green et al., *supra* note 35.

of observers to quickly detect and identify specific objects.[44] Figure 3 illustrates the perceptual consequence of visual crowding. In this image, letters are scaled, as in Figure 2B, to compensate for loss of visual acuity. In this case, however, the letters are "crowded" together in the central rings. As a consequence, the ability to detect and recognize specific letters is impaired. This effect suggests that a crime committed in a visually complex scene, such as a sporting event or a crowded street, will place severe and predictable limits on the ability of a witness to accurately perceive the facial features of a perpetrator.

Vision is also limited by image distortion (foreshortening) associated with angle of view. The retinal pattern generated by a face viewed directly from the front differs considerably—with changes in aspect ratio and relative placement of facial features—from that generated by a face viewed from an oblique side angle. Viewing a face from an angle above or below center (if the criminal were standing over you, or below you on the stairs) also yields retinal distortions

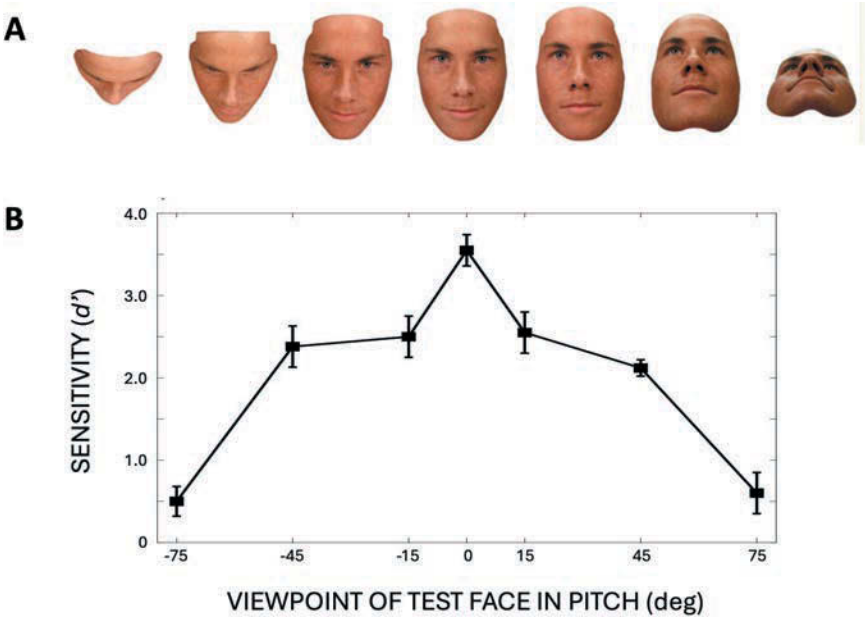
Figure 3. Visual "crowding" impairs object recognition.

Source: Used with permission of Elsevier Science & Technology Journals, from Anstis, *supra* note 40, at 589-592 (permission conveyed through Copyright Clearance Ctr., Inc.).

44. Dennis M. Levi, *Crowding—An Essential Bottleneck For Object Recognition: A Mini-Review*, 48 Vision Rsch. 635 (2008), https://doi.org/10.1016/j.visres.2007.12.009; David Whitney & Dennis M. Levi, *Visual Crowding: A Fundamental Limit on Conscious Perception and Object Recognition*, 15 Trends Cognitive Sci. 160 (2011), https://doi.org/10.1016/j.tics.2011.02.005; Ryan V. Ringer et al., *Investigating Visual Crowding of Objects in Complex Real-World Scenes*, 12 i-Perception (2021), https://doi.org/10.1177/2041669521994150.

(foreshortening) of facial features, as shown in Panel A of Figure 4.[45] Panel B of Figure 4 plots the ability of human observers to discriminate (*d'*) a frontal view (0 degrees) from the same or different face tilted up or down. Discriminability declined markedly with tilt angle. (The seven data points in Panel B correspond to the seven view angles in Panel A.) These effects are largely overcome when viewing familiar people because we have extensive prior knowledge of appearance that is independent of view angle. When viewing unfamiliar faces under time-limited conditions, however, this image distortion can severely impair subsequent recognition.[46]

Figure 4. Viewing angle affects face recognition.



Note: Panel A was created by the authors using parts extracted from Figure 1 on page 3 of source. Panel B was created by the authors by modifying Figure 2, Panel A, on page 5 of source.
Source: Adapted from Simone K. Favelle & Stephen Palmisano, *The Face Inversion Effect Following Pitch and Yaw Rotations: Investigating the Boundaries of Holistic Processing*, 3 Frontiers Psych. 563 (2012), https://doi.org/10.3389/fpsyg.2012.00563. Copyright © 2012 Favelle and Palmisano. This is an open-access article distributed under the terms of the Creative Commons Attribution License.

45. Simone K. Favelle & Stephen Palmisano, *The Face Inversion Effect Following Pitch and Yaw Rotations: Investigating the Boundaries of Holistic Processing*, 3 Frontiers Psych. 563 (2012), https://doi.org/10.3389/fpsyg.2012.00563.

46. In the extreme, unconventional views of a common object, such as a bucket viewed from above, make recognition nearly impossible in the absence of context.

Visual attention

William James, the great nineteenth-century experimental psychologist, famously articulated a subjective phenomenon familiar to us all:

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.[47]

By this rule, attention, which is required for awareness, is a limited resource. Thus, only a small fraction of the information falling on the retina reaches awareness or is used by an observer for recognition, action, or storage in memory. Allocation of selective attention is an active process that can be directed by external factors—visual attributes with high salience, such as a bright light or an unfamiliar object—or internal control.[48] For example, if you are searching for your phone, you may explicitly direct your attention to the countertop where it was last seen.

Attended image content is transiently enhanced—equivalent to turning up the volume of specific sensory signals—to increase the fidelity of perceptual experience. Image content not falling within the focus of attention is processed with less fidelity.[49] In some cases, unattended content is effectively invisible: It does not reach awareness, it is not perceived, and it is not available for use in guiding decisions or actions, or for storage in memory.[50]

Different pieces of visual information compete dynamically for attentional selection based on their changing physical attributes, locations in space, and relevance to the observer's needs and behavioral goals.[51] An eyewitness must select what to attend to, often within a short window of time, without advance warning, in the presence of many novel objects and events, and under the confounding influences of anxiety and fear. Reductions in efficiency are common under such conditions, causing sensitivity to unattended items to be markedly reduced when there are many objects simultaneously competing for attention.[52]

47. William James, The Principles of Psychology (Henry Holt ed., 1890).

48. Michael I. Posner, *Orienting of Attention*, 32 Q. J. Experimental Psych. 3 (1980), https://doi.org/10.1080/00335558008248231.

49. *Id.*

50. Arien Mack & Irvin Rock, Inattentional Blindness (1998).

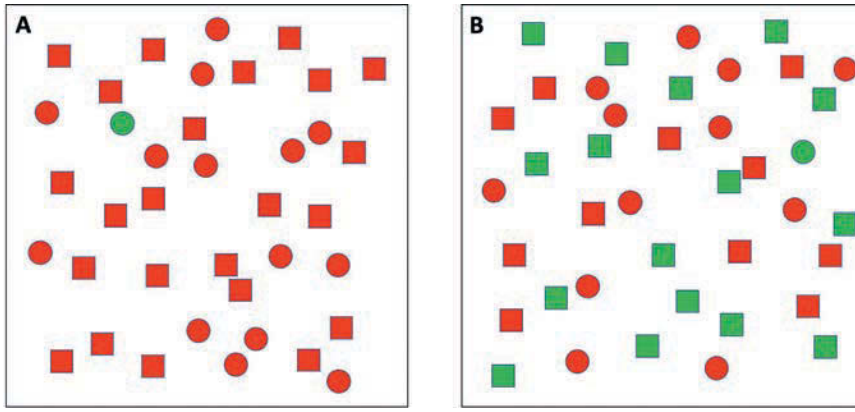
51. Harold Pashler, James C. Johnston & Eric Ruthruff, *Attention and Performance*, 52 Ann. Rev. Psych. 629 (2001), https://doi.org/10.1146/annurev.psych.52.1.629; Robert Desimone & John Duncan, *Neural Mechanisms of Selective Visual Attention*, 18 Ann. Rev. Neuroscience 193 (1995), https://doi.org/10.1146/annurev.ne.18.030195.001205.

52. James, *supra* note 47.

A simple demonstration of the complexity of attentive effects on awareness comes from the well-studied phenomenon of "visual search."[53] Consider the two images in Figure 5. In panel A, the green "target" stimulus immediately "pops out" perceptually from the field of red objects and commands attention, simply because it is distinguished by a difference in color; shape varies but is irrelevant to the task. By contrast, the target stimulus in panel B is distinguished by a unique *conjunction* of color and shape (green and circular); detection of this target requires an active attentional search that proceeds serially through the objects in the display.[54] What this means, of course, is that during time-limited viewing of an unfolding crime, a witness's attention will be automatically drawn to objects that are readily distinguished by a unique visual attribute—the woman in the red dress. Objects that are defined by unique conjunctions of features—the man on the bus in the gray sweater, surrounded by a field of men in gray coats and blue sweaters—will be harder to attend and, perhaps most importantly, harder to remember because of combinatorial complexity.

The effect of feature conjunctions on the perceptual experiences (also called percepts) that we commit to memory is illustrated by the image in Figure 6.[55] Look briefly at this image and then look away. Ask yourself whether the image contained a green letter *A*, a *B* surrounded by a diamond, or a red *D* in a circle. The red *D* in a circle is indeed present, but the *A* is blue and the *B* is surrounded

Figure 5. Focus of attention affects recognition of more complex objects.



53. Anne M. Treisman & Garry Gelade, *A Feature-Integration Theory of Attention*, 12 Cognitive Psych. 97 (1980), https://doi.org/10.1016/0010-0285(80)90005-5.

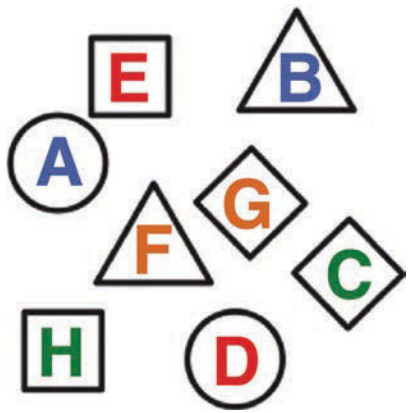
54. This search process has been described metaphorically as scanning the visual display using a "spotlight" of attention, which brings features in and out of awareness. *Id.*

55. This demonstration appears in Jeremy M. Wolfe, *Visual Search*, *in* Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (John T. Wixted ed., 2022), https://doi.org/10.1002/9781119170174.epcn213.

by a triangle. Your inability to accurately report what you have seen in this demonstration reflects the visual system's propensity to render "illusory conjunctions" of features.[56] The individual feature parameters—color, surrounding shape, and letter—are physically bound together in the image, but they are phenomenally unbound by vision, such that they easily recombine perceptually in forms that do not exist in the image. Once again, the implications for accurate eyewitness identification are significant: The person on the train platform who pulled the trigger had a mustache, brown hair, and light skin, but you may easily confuse these features with those of other people present in the scene, such that you describe him as possessing a goatee, blond hair, and light skin.

Another adverse consequence of attentional selection is that competing elements in a scene can hijack the attentional focus. The technique of misdirection—one of the essential tools of performance magic—directs attention to uninformative image content and exploits the invisibility of unattended features.[57] The well-studied phenomenon of inattentional blindness is an example of misdirection,

Figure 6. Errors of perception are committed to memory because of "illusory conjunctions" of features in a visual image.



Source: Used with permission of John Wiley & Sons, Inc., from Jeremy M. Wolfe, *Visual Search*, in Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (John T. Wixted ed., 2022), https://doi.org/10.1002/9781119170174.epcn213.

56. Anne Treisman & Hilary Schmidt, *Illusory Conjunctions in the Perception of Objects*, 14 Cognitive Psych. 107 (1982), https://doi.org/10.1016/0010-0285(82)90006-8.

57. Gustav Kuhn & Luis M. Martinez, *Misdirection–Past, Present, and the Future*, 5 Frontiers Hum. Neuroscience 172 (2012), https://doi.org/10.3389/fnhum.2011.00172; Stephen Macknik & Susana Martinez-Conde with Sandra Blakeslee, Sleights of Mind: What the Neuroscience of Magic Reveals About Our Everyday Deceptions (1st ed. 2010).

which occurs readily in real-world situations.[58] In this case, attention directed to one behaviorally significant property of a visual scene precludes awareness of other potentially important features.[59] As a result, an individual can be surprisingly unaware of surreptitious changes to the physical appearance of an unfamiliar person while the two are engaged in conversation.[60] This finding suggests that events that briefly divert attention have the potential to markedly impair the accuracy of eyewitness identifications.[61]

Attentional hijacking is particularly characteristic of stimuli that elicit strong emotional responses. Visual stimuli that trigger fear responses act as powerful external cues that command attention. While this potentiates sensitivity to those stimuli, at the expense of lost sensitivity to others, it is often the case that the attended emotional stimuli are not the ones with relevant informational content. The so-called weapon focus effect (reviewed below) is a real-world case in point for eyewitness identification, in which attention is compellingly drawn to emotionally laden stimuli, such as a gun or a knife, at the expense of acquiring greater visual information about the face of the perpetrator.[62] Similarly, viewing a crime with multiple perpetrators (see section titled "Multiple Perpetrators" below) divides attentional resources such that the fidelity of perceptual information about each face is limited.[63]

58. Mack & Rock, *supra* note 50; Ulric Neisser & Robert Becklen, *Selective Looking: Attending to Visually Specified Events*, 7 Cognitive Psych. 480 (1975), https://doi.org/10.1016/0010-0285(75)90019-5; Daniel J. Simons, *Attentional Capture and Inattentional Blindness*, 4 Trends Cognitive Scis. 147 (2000), https://doi.org/10.1016/s1364-6613(00)01455-8.

59. For a dramatic demonstration of this effect, produced by Daniel Simons and Christopher Chabris, see Daniel J. Simons, *Selective Attention Test*, http://tinyurl.com/inattentional-blindness; Daniel J. Simons & Christopher F. Chabris, *Gorillas in our Midst: Sustained Inattentional Blindness for Dynamic Events*, 28 Perception 1059 (1999), https://doi.org/10.1068/p281059.

60. Daniel J. Simons & Daniel T. Levin, *Failure to Detect Changes to People During a Real-World Interaction*, 5 Psychonomic Bull. & Rev. 644 (1998), https://doi.org/10.3758/bf03208840.

61. Graham Davies & Sarah Hine, *Change Blindness and Eyewitness Testimony*, 141 J. Psych. 423 (2007), https://doi.org/10.3200/jrlp.141.4.423-434.

62. Thomas H. Kramer, Robert Buckhout & Paul Eugenio, *Weapon Focus, Arousal, and Eyewitness Memory: Attention Must be Paid,* 14 Law & Hum. Behav. 167 (1990), https://doi.org/10.1007/bf01062971; Kerri L. Pickel, Stephen J. Ross & Ronald S. Truelove, *Do Weapons Automatically Capture Attention?*, 20 Applied Cognitive Psych. 871 (2006), https://doi.org/10.1002/acp.1235; Elizabeth F. Loftus, Geoffrey R. Loftus & Jane Messo, *Some Facts About "Weapon Focus,"* 11 Law & Hum. Behav. 55 (1987), https://doi.org/10.1007/bf01044839.

63. Brian R. Clifford & Clive R. Hollin, *Effects of the Type of Incident and the Number of Perpetrators on Eyewitness Memory*, 66 J. Applied Psych. 364 (1981), https://doi.org/10.1037//0021-9010.66.3.364; Zoe J. Hobson & Rachel Wilcock, *Eyewitness Identification of Multiple Perpetrators*, 13 Int'l J. Police Sci. & Mgmt. 286 (2011), https://doi.org/10.1350/ijps.2011.13.4.253; Robert F. Lockamyeir et al., *One Perpetrator, Two Perpetrators: The Effect of Multiple Perpetrators on Eyewitness Identification*, 35 Applied Cognitive Psych. 1206 (2021), https://doi.org/10.1002/acp.3853; Alicia Nortje, Colin G Tredoux & Annelies Vredeveldt, *Eyewitness Identification of Multiple Perpetrators*, 33 S. Afr. J. Crim. Just. 348 (2020), https://perma.cc/CN2K-JHSY.

Categorical perception

The precision of object recognition can easily be derailed because visual perception is categorical.[64] Although the objects of our experience vary extensively and uniquely along multiple sensory dimensions, we lump them into perceptual categories based on prior associations, many of which stem from common functions, physical properties, meanings, or emotional valence.[65] Many behavioral and cognitive processes are greatly simplified by treating all members of a category as the same, despite their physical differences. It rarely matters, for example, whether the pear we choose to eat is dappled on one side or irregular in shape, nor does the text font bear greatly on our ability to read. Evidence indicates that the structure of object memory is also categorical, suggesting that perceived objects are encoded in memory as a category type, often without specific detail, such that fine details needed for subsequent individuation are lost.[66]

One of the functional corollaries of categorical perception is that observers are far better at discriminating between objects from different categories than objects from the same category. It is easier, for example, to discriminate between yellow and orange than it is to discriminate different shades of orange, even when the wavelength differences are the same.[67] Another functional corollary is that it is possible to acquire better discriminative ability—perceptual expertise—that is sharply delimited by category boundaries.[68] Through exposure and practice, we can become experts at discriminating objects within a category, such as types of sports cars, x-radiographs, and tennis shoes, that each give us little benefit when judging objects outside of that category.

Not surprisingly, a well-known phenomenon of categorical expertise is manifested as better performance on a task that involves discriminating between faces of one's own race relative to faces of a different race. This cross-race effect

64. Robert L. Goldstone, *Influences of Categorization on Perceptual Discrimination*, 123 J. Exper. Psych.: Gen. 178 (1994), https://doi.org/10.1037//0096-3445.123.2.178.

65. Robert R. L. Goldstone & Andrew T. Hendrickson, *Categorical Perception*, 1 Wiley Interdisc. Revs.: Cognitive Sci. 69 (2010), https://doi.org/10.1002/wcs.26.

66. Robert L. Goldstone, Yvonne Lippa & Richard M. Shiffrin, *Altering Object Representations Through Category Learning*, 78 Cognition 27 (2001), https://doi.org/10.1016/s0010-0277(00)00099-8.

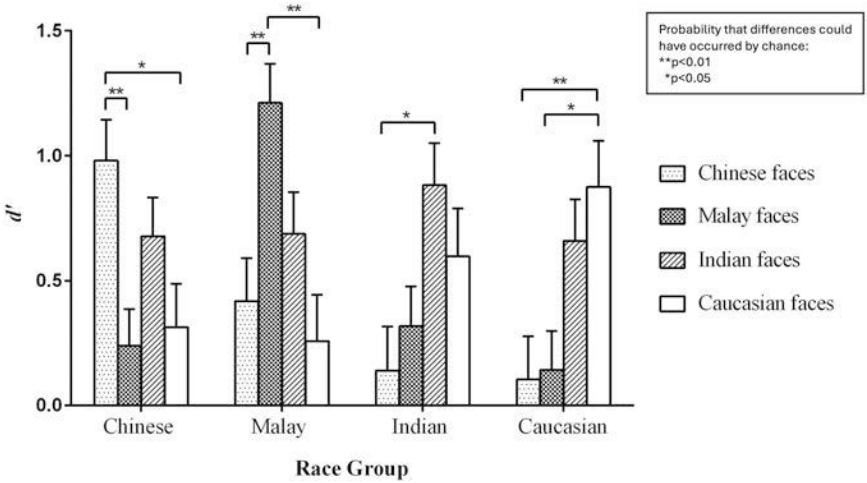
67. W. D. Wright & F. H. G. Pitt, *Hue-Discrimination in Normal Colour-Vision*, 46 Proceedings Physical Soc'y 459 (1934), https://doi.org/10.1088/0959-5309/46/3/317.

68. Eleanor J. Gibson, Principles of Perceptual Learning and Development (1969); Geoffrey R. Norman et al., *The Correlation of Feature Identification and Category Judgments in Diagnostic Radiology*, 20 Memory & Cognition 344 (1992), https://doi.org/10.3758/bf03210919; Rosalynn M. Peron & Gary L. Allen, *Attempts To Train Novices For Beer Flavor Discrimination: A Matter of Taste*, 115 J. Gen. Psych. 403 (1988), https://doi.org/10.1080/00221309.1988.9710577; Irving Biederman & Margaret M. Shiffrar, *Sexing Day-Old Chicks: A Case Study and Expert Systems Analysis of a Difficult Perceptual-Learning Task*, 13 J. Exper. Psych.: Learning, Memory & Cognition 640 (1987), https://doi.org/10.1037//0278-7393.13.4.640.

is pronounced and present in a variety of contexts.[69] In Figure 7, the y-axis plots a measure of face discriminability (*d'*) for faces of four different races. The x-axis identifies the race of the human observers performing the face discrimination task. Bar texture identifies the race of the faces being discriminated. Regardless of their own race, all observers are best at discriminating between faces of their own race. (Error bars in Figure 7 represent standard errors of the mean.) The differential perceptual expertise that underlies the cross-race effect is a natural consequence of perceptual learning from exposure biases (e.g., most children grow up surrounded by family and friends of the same race),[70] though there may be some consequence of in-group versus out-group social biasing effects.[71]

Category-specific perceptual expertise is an operating characteristic of biological sensory systems that affords the ability to gain proficiency in specialized domains, such as medical diagnosis of skin lesions, or the ability to quickly

Figure 7. Cross-race effect.



Source: Adapted from Wong et al., *supra* note 69, at 208. Copyright © 2020 Wong, Stephen and Keeble. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

69. Rankin W. McGugin et al., *Race-Specific Perceptual Discrimination Improvement Following Short Individuation Training with Faces*, 35 Cognitive Sci. 330–47 (2011), https://doi.org/10.1111/j.1551-6709.2010.01148.x; Hoo Heat Wong et al., *The Own-Race Bias for Face Recognition in a Multiracial Society*, 11 Frontiers Psych. 208 (2020), https://doi.org/10.3389/fpsyg.2020.00208.

70. McGugin et al., *supra* note 69.

71. Eric Hehman, Eric W. Mania & Samuel L. Gaertner, *Where the Division Lies: Common Ingroup Identity Moderates the Cross-Race Facial-Recognition Effect*, 46 J. Exper. Soc. Psych. 445 (2010), https://doi.org/10.1016/j.jesp.2009.11.008.

identify the gender of a baby chick.[72] But this same category specificity has obvious potential to yield inequity of recognition decisions under conditions of uncertainty in the real world.

The Constructive Nature of Visual Experience

Any effort to assess the validity of what one says they saw must come to grips with the fact that visual perception is, by its very nature and for very good reasons, not always an accurate reproduction of the world. Vision does not operate like a camera. Rather, visual perception is a *construct* of what we predict the world to be. To understand this constructive mandate, consider the optical and biological processes that underlie visual perception: Light reflected off of objects in the world around us is refracted by the lens in the anterior portion of the eye and projected as a two-dimensional (2D) patterned image onto the back surface of the eye, known as the retina. The goal of vision is to infer, from these 2D retinal patterns, the structure and meaning of objects and events in the three-dimensional (3D) world around us. This inferential problem is constrained, however, by the fact that the projected image does not contain—indeed *cannot* contain—all of the information needed to accurately determine the real-world cause of the image.[73] To put it bluntly, there is an infinite set of real-world environments that can give rise to any specific retinal image.

Visual perception overcomes much uncertainty about the cause of sensation through reference to prior knowledge or experience. From these priors, or "biases," the human brain fills in perceptual gaps with what is most likely to be out there. To illustrate this effect, consider the abstract visual pattern that appears in Figure 8. This pattern elicits a high degree of perceptual uncertainty; most people struggle to identify what it is. Figure 9 provides an additional bit of information that instills a bias. When returning to Figure 8, that bias enables the observer to improvise perceptual structure and meaning where formerly there was none.

Biases born from experience thus play a ubiquitous and essential role in perception. This concept is illustrated schematically in Figure 10. Information about the world gathered from sensory input, and information previously stored in memory, normally collaborate to yield coherent percepts. Thus, under most conditions, what we actually see falls somewhere along a continuum of mixtures of sensation and memory—a condition that facilitates behavioral interaction with the world under conditions of sensory uncertainty. The two ends of this continuum, however, are problematic states. At one end, pure stimulus is the

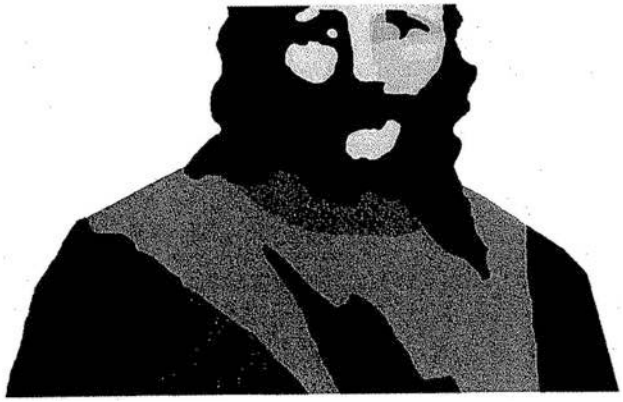
72. Goldstone, *supra* note 64; Biederman & Shiffrar, *supra* note 68.

73. This is a classic "inverse problem" in that there is not sufficient information in the retinal image to uniquely infer the values of the parameters that gave rise to it.

Figure 8. Perceptual uncertainty in response to visual "noise."

Source: Paul B. Porter, *Another Puzzle-Picture*, 67 Am. J. Psych. 550–51 (1954), https://doi.org/10.2307/1417952. Copyright 1954 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.

Figure 9. Clear image will resolve uncertainty in Figure 8.



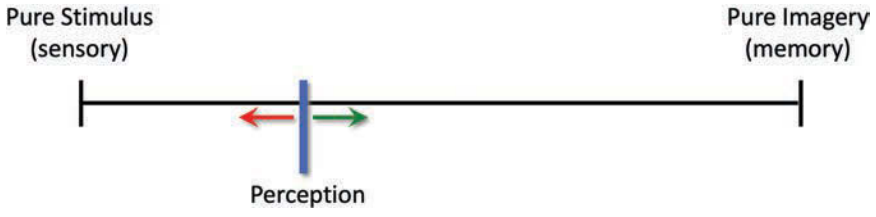
Source: Paul B. Porter, *Another Puzzle-Picture*, 67 Am. J. Psych. 550–51 (1954), https://doi.org/10.2307/1417952. Copyright 1954 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.

condition in which a meaningful interpretation is impossible. An observer's initial viewing of Figure 8 leads to that state. This type of perceptual failure is rare in adult visual experience, but it is debilitating and potentially dangerous when it occurs, because the observer is confronted with a stimulus for which rational decisions and actions are impossible. At the other end of the continuum, pure imagery is the hallucinatory condition in which perceptual experience is detached from all sensory input, which can be entertaining or terrifying, but not at all adaptive. There are thus selective pressures for perceptual improvisation, through integration of sensation and memory.

Though evolution has thus afforded us the ability to employ prior knowledge to perceive an unambiguous world under conditions of uncertainty, there is a well-known hazard associated with this process: Our priors are sometimes wrong. In one of the most provocative demonstrations of this phenomenon, Jerome Bruner and Leo Postman used "trick" playing cards to demonstrate a persistent biasing influence of prior knowledge on perception.[74] The trick cards were created by altering the color of a given suit—a red eight of spades, for example. Observers were shown a series of cards with brief presentations; some cards were trick and the others normal. With surprising frequency, observers failed to identify the trick cards and instead reported them as normal, thus demonstrating that strongly learned associations are capable of sharply, persistently, and unconsciously biasing visual perception. The heightened uncertainty associated with rapidly unfolding events of a crime is similarly ripe for unconscious introduction of biases held by the observer.

To make matters worse, perceptual inaccuracy born from uncertainty and bias is often associated with overconfidence—a special form of bias in which the observer implicitly rates the certainty of the experience as greater than it warrants. This phenomenon was pronounced in the "trick card" experiment cited above: Upon questioning about what they had perceived, the observers staunchly

Figure 10. Perceptual state falls on a continuum between pure stimulus and pure imagery.



74. Jerome S. Bruner & Leo Postman, *On the Perception of Incongruity: A Paradigm*, 18 J. Personality 206 (1949), https://doi.org/10.1111/j.1467-6494.1949.tb01241.x.

defended their mistaken identifications. Both eyewitness experiments conducted in the laboratory[75] and studies of actual criminal cases[76] demonstrate that while confidence may be low at the time of the initial lineup identification, confidence grows with reinforcement and with receipt of what appears as independent corroborating information—someone else saw something similar—which may drive the observer's certainty in line with a larger story (see section below titled "Confidence").[77]

In summary, the sense of vision is a biological information-processing device. As with any such device, the information gained is plagued by uncertainty, bias, and overconfidence. Uncertainty is a natural consequence of the incompleteness and ambiguity of sensory information. Biases drawn from memory fill in the gaps to ensure a seamless perceptual experience. Confidence in what was perceived grows as those experiences are reinforced by other sources of information. The significance of this perspective for eyewitness testimony is profound, as it helps us to realize that the accuracy of information about the environment—the face of a criminal perpetrator, for example—gained through vision is necessarily, and often sharply, limited by ambiguity and noise. It naturally follows that having an opportunity "to view the criminal at the time" (per *Manson*) may not be sufficient for assessment of the accuracy of eyewitness testimony, because what was viewed is not necessarily what was perceived and remembered.

75. Gary L. Wells, Tamara J. Ferguson & R. C. Lindsay, *The Tractability of Eyewitness Confidence and Its Implications for Triers of Fact*, 66 J. Applied Psych. 688 (1981), https://doi.org/10.1037//0021-9010.66.6.688; Melissa Paiva et al., *Influence of Confidence Inflation and Explanations for Changes in Confidence on Evaluations of Eyewitness Identification Accuracy*, 16 Legal & Criminological Psych. 266 (2011), https://doi.org/10.1348/135532510x503340; Siegfried Ludwig Sporer et al., *Choosing, Confidence, and Accuracy: A Meta-Analysis of the Confidence-Accuracy Relation in Eyewitness Identification Studies*, 118 Psych. Bull. 315 (1995), https://doi.org/10.1037//0033-2909.118.3.315; Kenneth A. Deffenbacher, *Eyewitness Accuracy and Confidence: Can We Infer Anything About Their Relationship?*, 4 Law & Hum. Behav. 243 (1980), https://doi.org/10.1007/bf01040617; Kenneth A. Deffenbacher & William Sturgill, *Some Predictors of Eyewitness Memory Accuracy*, *Practical Aspects of Memory*, 4 Law & Hum. Behav. 219 (1980), https://psycnet.apa.org/doi/10.1007/BF01040617; Mark Tippens Reinitz et al., *Confidence–Accuracy Relations for Faces and Scenes: Roles of Features and Familiarity*, 19 Psychonomic Bull. & Rev. 1085 (2012), https://doi.org/10.3758/s13423-012-0308-9; Robert K. Bothwell, Kenneth A. Deffenbacher & John C. Brigham, *Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited*, 72 J. Applied Psych. 691 (1987), https://doi.org/10.1037//0021-9010.72.4.691; Amy Bradfield Douglass & Eric E. Jones, *Confidence Inflation in Eyewitnesses: Seeing Is Not Believing*, 18 Legal & Criminological Psych. 152 (2020), https://doi.org/10.1111/j.2044-8333.2011.02031.x; Laura Smalarz & Gary L. Wells, *Do Multiple Doses of Feedback Have Cumulative Effects on Eyewitness Confidence?*, 9 J. Applied Rsch. Memory & Cognition 508 (2020), https://doi.org/10.1037/h0101857; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications*, 20 Psych., Pub. Pol'y. & L. 1 (2014), https://doi.org/10.1037/law0000001; Amy Bradfield Douglass & Nancy Steblay, *Memory Distortion in Eyewitnesses: A Meta-Analysis of the Post-Identification Feedback Effect*, 20 Applied Cognitive Psych. 859 (2006), https://doi.org/10.1002/acp.1237.

76. Garrett, *supra* note 2.

77. Daniel Kahneman & Amos Tversky, *On the Psychology of Prediction*, 80 Psych. Rev. 237 (1973), https://doi.org/10.1037/h0034747.

How Memory Works

Functional Processes of Memory

Visual perceptual experiences are commonly stored as declarative memories, which are accessed and expressed as knowledge about the world. Declarative memories are of two types: Semantic memories are of facts and concepts, whereas episodic memories are of events, such as those witnessed during a crime or a walk in the park. Declarative memories are processed in three stages—encoding, storage, and retrieval—which refer to placement of items in memory, maintenance therein, and subsequent access to the stored information.

In the following sections, we briefly review memory encoding, storage, and retrieval, with emphasis on the limits of these processes as they pertain to eyewitness identification. We also discuss the phenomenon of "recognition memory," which refers to the specific type of memory retrieval in which a stimulus (e.g., a face) is used to make an identification.

Memory Encoding

Encoding is the process through which perceived objects and events are placed into storage. This process occurs in two temporal stages, distinguished by the quantity of information stored, the duration of storage, and the susceptibility to interference.[78] Short-term or working memory is the conscious content of recent perceptual experiences, or information recently recalled from long-term storage. Short-term memory is of limited duration and capacity[79] and is labile, decaying quickly and easily disrupted by other perceptual or cognitive processes.[80]

Through cellular and molecular events that play out over time, the contents of short-term memories may be encoded and consolidated into long-term memory,[81] which is more enduring (albeit evolving with ongoing experience), and of greater capacity. Memories are particularly labile during this encoding process. The contents can be disrupted by interference and biased by prior

78. R. C. Atkinson & R. M. Shiffrin, *Human Memory: A Proposed System and Its Control Processes*, *in* 2 The Psychology of Learning and Motivation 89 (1968), https://doi.org/10.1016/s0079-7421(08)60422-3; James, *supra* note 47; Alan Baddeley, *Working Memory: Looking Back and Looking Forward*, 4 Nature Revs. Neurosci. 829 (2003), https://doi.org/10.1038/nrn1201; Alan Baddeley, Working Memory (1986).

79. George A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, 63 Psych. Rev. 81 (1956), https://doi.org/10.1037/h0043158.

80. John Jonides et al., *The Mind and Brain of Short-Term Memory*, 59 Ann. Rev. Psych. 193 (2008), https://doi.org/10.1146/annurev.psych.59.103006.093615.

81. Eric Kandel & Larry Squire, Memory: From Mind to Molecules (1999).

knowledge, expectations, or beliefs, resulting in a distorted representation of experience. Short-term memories of events that happened early in a witnessed proceeding may simply be forgotten with the passage of time or compromised by attention directed to subsequent emotional events or cognitive and behavioral demands (e.g., anxiety, fear, the need to escape). In such cases, the compromised information may never be consolidated fully into long-term storage, or that storage may contain distorted content.[82] At the same time, the quality of encoding of stimuli that are attended is commonly enhanced by highly emotional content.[83]

Memory Storage

Once memories have been encoded, they may be retained indefinitely in long-term storage. That stored information is not immutable, however, like information locked in a vault. On the contrary, stored memories fade with the passage of time, and they are commonly modified—often without awareness—as new knowledge and experiences are acquired. The quantitative features of these effects can provide insight into the validity of eyewitness reports.

How memory goes missing: The "forgetting function"

Much of the plasticity that plagues memory storage is the simple loss of information, which we call forgetting. The famous Ebbinghaus "forgetting function," reproduced in Figure 11, illustrates that under the conditions used in Ebbinghaus's nineteenth-century experiment (recall of nonsense syllables), two-thirds of the experienced content was forgotten after 24 hours.[84] The rate of forgetting

82. James L. McGaugh, *Memory—a Century of Consolidation*, 287 Sci. 248 (2000), https://doi.org/10.1126/science.287.5451.248; James L. McGaugh & Benno Roozendaal, *Role of Adrenal Stress Hormones in Forming Lasting Memories in the Brain*, 12 Current Op. Neurobiology 205 (2002), https://doi.org/10.1016/s0959-4388(02)00306-9.

83. Kevin N. Ochsner, *Are Affective Events Richly Recollected or Simply Familiar? The Experience and Process of Recognizing Feelings Past*, 129 J. Exper. Psych.: Gen. 242 (2000), https://doi.org/10.1037//0096-3445.129.2.242; Deborah Talmi et al., *Immediate Memory Consequences of the Effect of Emotion on Attention to Pictures*, 15 Learning & Memory 172 (2008), https://doi.org/10.1101/lm.722908; Elizabeth A. Kensinger & Daniel L. Schacter, *Neural Processes Supporting Young and Older Adults' Emotional Memories*, 7 J. Cognitive Neurosci. 1 (2008), https://doi.org/10.1162/jocn.2008.20080; Elizabeth A. Phelps, *Emotion and Cognition: Insights from Studies of the Human Amygdala*, 57 Ann. Rev. Psych. 27 (2006), https://doi.org/10.1146/annurev.psych.56.091103.070234.

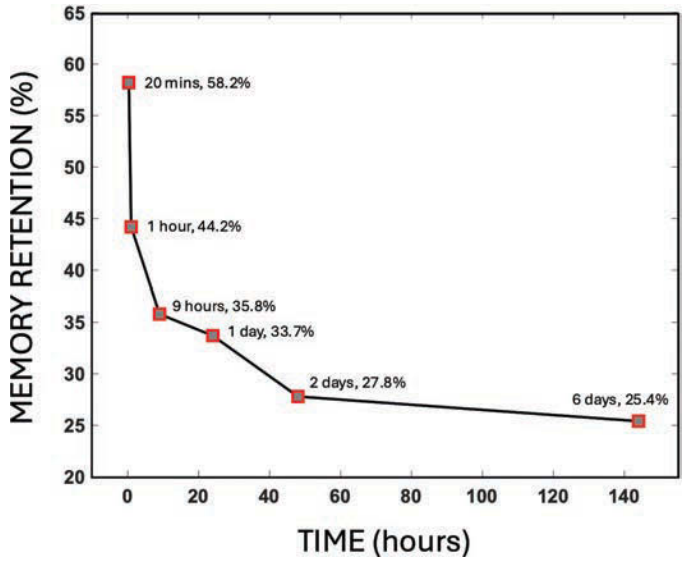
84. Hermann Ebbinghaus, Memory: A Contribution to Experimental Psychology (Henry A. Ruger & Clara E. Bussenius trans., 1913) (1885).

has been studied in greater detail over the past 150 years. Effects depend upon the type of memory content and the circumstances under which it was encoded, but the time-dependent decline follows a power law: $m = Lt^{-D}e^{-It}$, where m is a measure of memory retained, L is the strength of memory at time 0, t is retention time in seconds, D is decay rate, and I is interference rate.[85] The modeled value of decay captures the rapid initial decline in memory retention, which is a consequence of failure to consolidate new information into long-term memory. As consolidation proceeds with the passage of time, there is a corresponding reduction of the rate of forgetting.

How memory goes wrong: Source memory failure and false memories

Memory loss is not simply due to decay. Memory mechanisms are inherently plastic throughout life, which means that content stored for the long term is labile in the face of new information. Without awareness, we regularly reconstruct,

Figure 11. Ebbinghaus forgetting function.



85. Wayne A. Wickelgren, *Single-Trace Fragility Theory of Memory Dynamics*, 2 Memory & Cognition 775 (1974), https://doi.org/10.3758/bf03198154; John T. Wixted & Ebbe B. Ebbesen, *On the Form of Forgetting*, 2 Psych. Sci. 409 (1991), https://doi.org/10.1111/j.1467-9280.1991.tb00175.x.

update, and distort the things we believe to be true.[86] Because new content can be added and the source of that content forgotten (source memory failure, in which we effectively forget how we know things),[87] an observer may attribute updated memories to the originally witnessed events—in some cases, substantially changing what they believe they have seen. An eyewitness might learn from the police that a potential suspect has a goatee and then attribute that knowledge to the witnessed events, which could have disastrous consequences for the ability of the eyewitness to accurately report what was seen. Opportunities for this interference naturally increase with the passage of time, countering ongoing consolidation of new memories. Newly incorporated information also need not be true to fact. Research on false memories shows that it is possible to plant fabricated content in memory, which leads to recall of things that were never experienced.[88]

The effect of emotion on memory storage

Memories of highly arousing emotional stimuli tend to be more enduring than memories of nonarousing stimuli.[89] Highly salient, unexpected, or arousing events—such as the *Challenger* space shuttle disaster, September 11, or episodes associated with a witnessed crime—are commonly stored strongly in memory, and their later retrieval is often associated with the subjective experience of high

86. Elizabeth F. Loftus & Geoffrey R. Loftus, *On the Permanence of Stored Information in the Human Brain*, 35 Am. Psych. 409 (1980), https://doi.org/10.1037//0003-066x.35.5.409.

87. D. Stephen Lindsay & Marcia K. Johnson, *Recognition Memory and Source Monitoring*, 29 Bull. Psychonomic Soc'y 203 (1991), https://doi.org/10.3758/bf03335235.

88. Elizabeth F. Loftus, *Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory*, 12 Learning & Memory 361 (2005), https://doi.org/10.1101/lm.94705; Elizabeth F. Loftus & Jacqueline E. Pickrell, *The Formation of False Memories*, 25 Psych. Annals 720 (1995), https://doi.org/10.3928/0048-5713-19951201-07; Marcia K Johnson & Carol L Raye, *False Memories and Confabulation*, 2 Trends Cognitive Scis. 137 (1998), https://doi.org/10.1016/s1364-6613(98)01152-8.

89. Lewis J. Kleinsmith & Stephen Kaplan, *Paired-Associate Learning as a Function of Arousal and Interpolated Interval*, 65 J. Exper. Psych. 190 (1963), https://doi.org/10.1037/h0040288; Michael W. Eysenck, *Arousal, Learning, and Memory*, 83 Psych. Bull. 389 (1976), https://doi.org/10.1037//0033-2909.83.3.389; Friderike Heuer & Daniel Reisberg, *Vivid Memories of Emotional Events: The Accuracy of Remembered Minutiae*, 18 Memory & Cognition 496 (1990), https://doi.org/10.3758/bf03198482; Tali Sharot & Elizabeth A. Phelps, *How Arousal Modulates Memory: Disentangling the Effects of Attention and Retention*, 4 Cognitive, Affective & Behav. Neurosci. 294 (2004), https://doi.org/10.3758/cabn.4.3.294; Elizabeth A. Kensinger, Rachel J. Garoff-Eaton & Daniel L. Schacter, *Memory for Specific Visual Details Can Be Enhanced by Negative Arousing Content*, 54 J. Memory & Language 99 (2006), https://doi.org/10.1016/j.jml.2005.05.005; Elizabeth A. Kensinger, *Remembering Emotional Experiences: The Contribution of Valence and Arousal*, 15 Revs. Neuroscis. 241 (2004), https://doi.org/10.1515/revneuro.2004.15.4.241.

vividness and a sense of reliving.[90] The stronger encoding and storage of emotional memories results from the engagement of a specialized system of stress hormones, which is triggered by arousing content and has potentiating effects on memory consolidation and storage.[91] Despite the vividness that characterizes retrieval of emotional memories, there are many indications that such memories are prone to errors.[92] Although emotional memories are often inaccurate in detail, one important corollary of their vividness is that they are frequently held with high confidence.[93] This breakdown of the relationship between accuracy and confidence can undermine eyewitness accounts.[94]

90. Gezinus Wolters & Joop J. Goudsmit, *Flashbulb and Event Memory of September 11, 2001: Consistency, Confidence and Age Effects*, 96 Psych. Reps. 605 (2005), https://doi.org/10.2466/pr0.96.3.605-619; Elizabeth A. Kensinger, Anne C. Krendl & Suzanne Corkin, *Memories of an Emotional and a Nonemotional Event: Effects of Aging and Delay Interval*, 32 Exper. Aging Rsch. 23 (2006), https://doi.org/10.1080/01902140500325031; Ulric Neisser & Nicole Harsch, *Phantom Flashbulbs: False Recollections of Hearing the News about* Challenger, *in* Affect and Accuracy in Recall: Studies of "Flashbulb" Memories 9 (Emory Symposia in Cognition, 1992), https://doi.org/10.1017/CBO9780511664069.003; K. S. LaBar & E. A. Phelps, *Arousal-Mediated Memory Consolidation: Role of the Medial Temporal Lobe in Humans*, 9 Psych. Sci. 490 (1998), https://doi.org/10.1111/1467-9280.00090.

91. James L. McGaugh, *Memory—A Century of Consolidation*, 287 Sci. 248 (2000), https://doi.org/10.1126/science.287.5451.248; James L McGaugh & Benno Roozendaal, *Role of Adrenal Stress Hormones in Forming Lasting Memories in the Brain*, 12 Current Op. Neurobiology 205 (2002), https://doi.org/10.1016/s0959-4388(02)00306-9.

92. Elizabeth A. Kensinger, *Remembering the Details: Effects of Emotion*, 1 Emotion Rev. 99 (2009), https://doi.org/10.1177/1754073908100432; Tali Sharot, Mauricio R Delgado & Elizabeth A Phelps, *How Emotion Enhances the Feeling of Remembering*, 7 Nature Neurosci. 1376 (2004), https://doi.org/10.1038/nn1353; H. Schmolck, E.A. Buffalo & L.R. Squire, *Memory Distortions Develop Over Time: Recollections of the O.J. Simpson Trial Verdict After 15 and 32 Months*, 11 Psych. Sci. 39 (2000), https://doi.org/10.1111/1467-9280.00212; Stephen R. Schmidt, *Autobiographical Memories for the September 11th Attacks: Reconstructive Errors and Emotional Impairment of Memory*, 32 Memory & Cognition 443 (2004), https://doi.org/10.3758/bf03195837; Tony W. Buchanan & Ralph Adolphs, *The Role of the Human Amygdala in Emotional Modulation of Long-Term Declarative Memory, in Emotional Cognition: From Brain to Behavior, in* Advances in Consciousness Research 9 (2002), https://doi.org/10.1075/aicr.44.02buc.

93. Ulrike Rimmele et al., *Emotion Enhances the Subjective Feeling of Remembering, Despite Lower Accuracy for Contextual Details*, 11 Emotion 553 (2011), https://doi.org/10.1037/a0024246; Kensinger, *supra* note 92; Neisser & Harsch, *supra* note 90; Elizabeth A. Phelps & Tali Sharot, *How (and Why) Emotion Enhances the Subjective Sense of Recollection*, 17 Current Directions Psych. Sci. 147 (2008) https://doi.org/10.1111/j.1467-8721.2008.00565.x.

94. Kate A. Houston et al., *The Emotional Eyewitness: The Effects of Emotion on Specific Aspects of Eyewitness Recall and Recognition Performance*, 13 Emotion 118 (2013), https://doi.org/10.1037/a0029220; Robin S. Edelstein et al., *Emotion and Eyewitness Memory*, *in* Memory and Emotion 308 (D. Reisberg & P. Hertel eds., 2004), https://doi.org/10.1093/acprof:oso/9780195158564.003.0010; Sven-Åke Christianson, *Emotional Stress and Eyewitness Memory: A Critical Review*, 112 Psych. Bull. 284 (1992), https://doi.org/10.1037//0033-2909.112.2.284.

Memory Retrieval

Memory retrieval refers to the process by which stored information is accessed and brought into consciousness, where it can be used to make decisions and guide actions. Retrieval of long-term declarative memories is commonly triggered through association with an external stimulus—the sound of the coffee grinder is a powerful retrieval cue for visual and olfactory memories of the morning espresso. A corollary of this association-based phenomenon is that memory retrieval is often context dependent; a memory may be more readily retrieved if the observer is in physical surroundings that are the same as or similar to those in which the original experiences took place,[95] because the surroundings provide additional cues to trigger memory retrieval.

Memory retrieval is affected by various sources of noise. Similarities of meaning or appearance between retrieval cues and items in memory can easily lead to retrieval of the wrong item, producing disruptive cognitive experiences—the sound of thunder may terrorize a veteran with retrieved memories of the battlefield—or false reports of what had been stored.[96] This is particularly a problem given the categorical nature of memory.[97] A related form of noise consists of intrusion errors, in which information known to be commonly associated with events of a general type becomes incorporated into the retrieved content of a specific memory. For example, because guns are often associated with robbery, an observer may readily and unwittingly

95. D. R. Godden & A. D. Baddeley, *Context Dependent Memory in Two Natural Environments: On Land and Underwater*, 66 Brit. J. Psych. 325 (1975), https://doi.org/10.1111/j.2044-8295.1975.tb01468.x; Stephen M. Smith & Edward Vela, *Environmental Context-Dependent Eyewitness Recognition*, 6 Applied Cognitive Psych. 125 (1992), https://doi.org/10.1002/acp.2350060204; Stephen M. Smith & Edward Vela, *Environmental Context-Dependent Memory: A Review and Meta-Analysis*, 8 Psychonomic Bull. Rev. 203 (2001), https://doi.org/10.3758/bf03196157; Endel Tulving & Donald M. Thomson, *Encoding Specificity and Retrieval Processes in Episodic Memory*, 80 Psych. Rev. 352 (1973), https://doi.org/10.1037/h0020071.

96. John R. Anderson, *A Spreading Activation Theory of Memory*, 22 J. Verbal Learning & Verbal Behav. 261 (1983), https://doi.org/10.1016/s0022-5371(83)90201-3; Allan M. Collins & Elizabeth F. Loftus, *A Spreading-Activation Theory of Semantic Processing*, 82 Psych. Rev. 407 (1975), https://doi.org/10.1037//0033-295x.82.6.407; Henry L. Roediger III, David A. Balota & Jason M. Watson, *Spreading Activation and Arousal of False Memories*, *in* The Nature of Remembering: Essays in Honor of Robert G. Crowder 95 (2001), https://doi.org/10.1037/10394-006; C. J. Brainerd & V. F. Reyna, The Science of False Memory (2005), https://doi.org/10.1093/acprof:oso/9780195154054.003.0002.

97. Endel Tulving, *Episodic and Semantic Memory: Where Should We Go From Here?*, 9 Behav. & Brain Scis. 573 (1986), https://doi.org/10.1017/s0140525x00047257; Martha J. Farah & James L. McClelland, *A Computational Model of Semantic Memory Impairment: Modality Specificity and Emergent Category Specificity*, 120 J. Exper. Psych.: Gen. 339 (1991), https://doi.org/10.1037//0096-3445.120.4.339; Glyn W. Humphreys & Emer M. E. Forde, *Hierarchies, Similarity, and Interactivity in Object Recognition: "Category-Specific" Neuropsychological Deficits*, 24 Behav. & Brain Scis. 453 (2001), https://doi.org/10.1017/s0140525x01004150.

incorporate a gun into the retrieved version of his or her memory of a witnessed robbery.

As noted above, memory retrieval is sometimes facilitated by being in the physical surroundings where the memorized events were experienced. A related phenomenon is state-dependent memory, in which retrieval accuracy is best if the individual's cognitive state at the time of retrieval matches the cognitive state at the time of encoding.[98] When memories have an emotional component, retrieval may be best when the individual is induced to a corresponding emotional state,[99] which is accomplished by verbally or physically placing the person in the same context.[100]

Recognition Memory

Recognition memory is the specific type of declarative memory retrieval in which an immediately present sensory stimulus elicits retrieval of the same stimulus stored following a prior encounter. The subjective determination of "sameness" is the key here, for recognition memory differs from other forms of retrieval (such as recalling a phone number or the recipe for lasagna) in that a comparison must be made between the retrieved memory and the sensory stimulus (the "cue") that elicited retrieval. This comparison process ultimately yields a categorical decision about whether the cue and the retrieved memory of a sensory stimulus are sufficiently similar (same source) or not (different source).

Recognition memory for faces differs greatly between familiar and unfamiliar faces. Because we often identify familiar faces with ease, most people tend to think they are generally very good at face recognition. However, the ability to recognize unfamiliar faces differs widely across individuals. At one extreme are those people, referred to as "super recognizers," who hardly forget a face. At the other end of the spectrum are "face-blind people," who have great difficulty

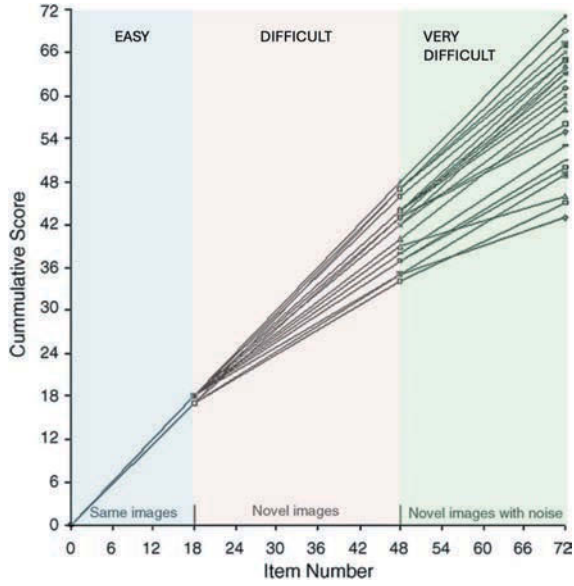
98. Donald W. Goodwin et al., *Alcohol and Recall: State-Dependent Effects in Man*, 163 Sci. 1358 (1969), https://doi.org/10.1126/science.163.3873.1358; Endel Tulving & Donald M. Thomson, *Encoding Specificity and Retrieval Processes in Episodic Memory*, 80 Psych. Rev. 352 (1973), https://doi.org/10.1037/h0020071; Edward Girden & Elmer Culler, *Conditioned Responses in Curarized Striate Muscle in Dogs*, 23 J. Compar. Psych. 261 (1937), https://doi.org/10.1037/h0058634; Donald A. Overton, *State Dependent or "Dissociated" Learning Produced with Pentobarbital*, 57 J. Compar. & Physiological Psych. 3 (1964), https://doi.org/10.1037/h0048023.

99. Pamela M. Kenealy, *Mood State-Dependent Retrieval: The Effects of Induced Mood on Memory Reconsidered*, 50 Q. J. Exper. Psych. 290 (1997), https://doi.org/10.1080/713755711; Penelope A. Lewis & Hugo D. Critchley, *Mood-Dependent Memory*, 7 Trends Cognitive Scis. 431 (2003), https://doi.org/10.1016/j.tics.2003.08.005; G. H. Bower, *Mood and Memory*, 36 Am. Psych. 129 (1981), https://psycnet.apa.org/doi/10.1037/0003-066X.36.2.129; Craik & Lockhart, *supra* note 22; Kensinger, *supra* note 92; Kenneth A. Leight & Henry C. Ellis, *Emotional Mood States, Strategies, and State-Dependency in Memory*, 20 J. Verbal Learning & Verbal Behav. 251 (1981), https://doi.org/10.1016/s0022-5371(81)90406-0.

100. Smith & Vela, *A Review and Meta-Analysis*, *supra* note 95.

recognizing even highly familiar faces. Evidence indicates that the ability to recognize an unfamiliar face varies continuously between these extremes across the human population.[101] Figure 12 shows the results of an experiment designed to demonstrate this variation. Observers performed a recognition memory task that involved matching faces of three difficulty levels, identified on the x-axis as "same" (easy), "novel" (difficult), and "novel with noise" (very difficult). The y-axis plots recognition performance ("cumulative score"), where perfect recognition corresponds to the case in which y = x. Each curve contains data from one of 25 subjects. For the easy category, performance was nearly equivalent across subjects and nearly perfect. For difficult faces, performance was more variable, with some subjects performing nearly perfectly and other subjects performing approximately 65% correct. For very difficult faces, performance was highly variable across subjects, with the best performers at nearly 100% correct and the

Figure 12. Natural variation in human face recognition ability.



Source: Used with permission of Elsevier Science & Technology Journals, adapted from Brad Duchaine & Ken Nakayama, *The Cambridge Face Memory Test: Results for Neurologically Intact Individuals and an Investigation of Its Validity Using Inverted Face Stimuli and Prosopagnosic Participants*, 44 Neuropsychologia, 576–85 (2006), https://doi.org/10.1016/j.neuropsychologia.2005.07.001 (permission conveyed through Copyright Clearance Ctr, Inc.).

101. Brad Duchaine & Ken Nakayama, *The Cambridge Face Memory Test: Results for Neurologically Intact Individuals and an Investigation of Its Validity Using Inverted Face Stimuli and Prosopagnosic.*

worst at less that 60% correct.[102] New efforts to assess where a given witness lies along this spectrum, summarized in the section below titled "Applied Eyewitness Studies in the Field," may prove diagnostic for interpreting the validity of identification testimony.[103]

How People Make Decisions

An eyewitness observer is manifestly a biological instrument for information measurement, storage, and classification. The foregoing discussions of vision and memory highlight functional capacities and operational limits of brain systems that serve the measurement and storage functions. The information measured and stored by an eyewitness is ultimately used to make a decision about identity, which is typically rendered as a binary choice ("He's the one," or "He's not the one") applied to each face in a lineup. The accuracy of that classification is an index of eyewitness performance. An understanding of the processes that underlie human decisions sheds light on where information processing works accurately or may go wrong.

Many human decisions involve comparisons of two items that have been measured by the senses. These items are often experienced as percepts derived from immediately present sensory stimuli: Which chair is bigger? Which light is brighter? Which suitcase is heavier? In some cases, the items being compared come from different informational sources, one from an immediate sensory stimulus and the other from memory of a stimulus previously experienced: Does it look like an oak tree? Does it taste like chicken? In both sensory-sensory (one sensory stimulus versus another simultaneously present sensory stimulus) and sensory-memory (a present sensory stimulus versus a remembered stimulus) comparisons, the results fall on a continuous cognitive scale of similarity. In all of these examples, however, the expected decision is not continuous; it is categorical, for example, "This one is brighter," or "No, it does not taste like chicken."

Human decisions thus depend upon two quantifiable elements. The first of these is the "decision variable," which, for the examples given, is the subjective measure of similarity. An observer's report of recognition is also influenced by the level of similarity that the observer finds acceptable. This level, known as the "decision criterion," is thus the second element that underlies human categorical decisions. Under some conditions, the sensory and memory measures being compared are so different that the threshold level for similarity is easy to assign.

102. *Id.*

103. *See* Jesse H. Grabman et al., *Predicting High Confidence Errors in Eyewitness Memory: The Role of Face Recognition Ability, Decision-Time, and Justifications*, 8 J. Applied Rsch. Memory & Cognition 233, 241 (2019), https://doi.org/10.1037/h0101835.supp.

There are, however, many similarity outcomes that are plausible under two different ground truth conditions (same versus different source), which make an observer's placement of the criterion for a categorical decision particularly problematic.

The decision criterion we actually apply under difficult conditions depends, in part, upon how we prioritize the outcomes. The simple desire to be informative to law enforcement, for example, may lead an eyewitness to lower their criterion for an identification decision. Estimating (or controlling) the observer's decision criterion is thus a critical step in efforts to judge the validity of an identification.

Applied Eyewitness Studies in the Laboratory

Overview and Significance

Coincident with Supreme Court rulings on eyewitness identification,[104] the 1970s saw a rapid growth of research on the validity of eyewitness testimony. In a field-defining review published in 1978,[105] Gary Wells developed a taxonomy of variables that would be expected to influence eyewitness performance, which unleashed a tide of scientific research on the effects of those variables in controlled laboratory studies designed to simulate real-world conditions.

The variables of interest, which reflect environmental and human factors, are particularly important because they are frequently at issue in criminal cases, and their known states may provide insight into the validity of eyewitness testimony. These variables fall into two classes, which Wells termed "estimator" and "system" variables.[106] Estimator variables characterize viewing conditions and the perceptual and cognitive states of the witness at the time of the crime. System variables, by contrast, influence identification accuracy after the crime has occurred and are variables that can be controlled by police.

The most valuable feature of applied laboratory studies of the impact of estimator and system variables is that "ground truth" is known. We thus know whether a given eyewitness made a correct or incorrect identification decision, which provides information about the predictive value of conditions associated with witnessing events (estimator variables) and identification events (system variables). The standard experimental paradigm involves exposing human

104. Neil v. Biggers, 409 U.S. 188 (1972); Manson v. Brathwaite, 432 U.S. 98, 114 (1977).

105. G. L. Wells, *Applied Eyewitness-Testimony Research: System Variables and Estimator Variables*, 36 J. Personality & Soc. Psych. 1546 (1978) https://doi.org/10.1037/0022-3514.36.12.1546.

106. *Id.*

subjects to mock crimes, in the form of staged events or videos. Subjects are then presented with a lineup and asked to report whether it contains a person they recognize from the witnessed crime. One can measure their accuracy, or whether they made a correct or incorrect choice, as well as other useful gauges of performance, such as confidence in their decision, and how quickly they arrive at that decision.

Although findings from many applied eyewitness studies have reinforced what basic research long ago revealed about the operating characteristics of vision and memory, particularly studies of the effects of viewing conditions and the cognitive state of the observer (estimator variables), these studies do so using a real-world context that highlights relevance to the courts. Applied research on other eyewitness variables, such as those associated with the type of lineup used for identification (system variables), has led to novel insights that can be used to improve eyewitness performance.

Performance Measures

The eyewitness identification problem studied under known-source laboratory conditions is a two-choice classification task, in which the witness subject must report whether a suspect is, or is not, the person they viewed at the crime scene. A few words about assessment of performance on this type of task are useful here.

Discrimination

The ability of an eyewitness to discriminate the perpetrator from innocent suspects is a core question in many laboratory studies, as it can be used to evaluate the impact of estimator and system variables on eyewitness performance. The standard measure of discriminability used for this purpose today is assessed as the frequency with which a witness identifies the perpetrator, relative to the frequency with which the witness identifies an innocent suspect, integrated across all decision criteria, ranging from circumspect to reckless.[107]

The knowledge gained from laboratory studies that employ this discriminability measure is an indispensable basis for policy decisions about the best lineup procedures. The level of discriminability associated with different lineup procedures can also be used to infer their relative effectiveness at yielding accurate testimony. This approach reveals nothing, however, about the probability

107. This has long been an effective tool used in basic research on the performance of biological systems for sensation and memory. *See* D. M. Green & J. A. Swets, Signal Detection Theory and Psychophysics (1966).

that a specific witness identification is correct, which is of course what the trier of fact really needs to know.[108]

Accuracy

Probability of correctness, or "accuracy," is distinct from discriminability and is defined in this context as the ratio of correct identifications of the perpetrator relative to all (correct or incorrect) identifications reported. Accuracy of a population of subjects is a simple thing to measure in laboratory studies because ground truth is known, and many recent studies have done so. Like discriminability, accuracy can be used to assess the diagnostic utility of evidence under specified conditions. The important question here is whether it is possible to predict accuracy in real criminal cases, where ground truth is not known. Accuracy prediction in this case must be based on other variables that have been shown to correlate with accuracy, such as DNA genotyping.

Estimator Variables

The states of estimator variables reflect viewing conditions and cognitive status of the witness. Although not controllable by the criminal justice system, these states can be quantified for a given instance and used to predict the fidelity of vision and memory. In what follows, we examine several estimator variables that are commonly associated with real criminal cases in order to illustrate the range of effects on performance as demonstrated in laboratory studies.

Weapon Focus

"Weapon focus" is a phenomenon in which a witness's focus of visual attention is directed away from the object that may be the most informative in the long run—the face of the perpetrator—and toward the object that is perceived as most immediately threatening to life.[109] As we have learned from basic research on vision (see section titled "Visual attention" above), attention is a limited resource that can be hijacked by compelling features in a visual scene. Just as a driver must focus attention on oncoming traffic at the expense of enjoying scenery along the road, a witness's life may depend upon the ability to keep attention focused on the perpetrator's

108. David Dunning & Liza Beth Stern, *Distinguishing Accurate from Inaccurate Eyewitness Identifications via Inquiries About Decision Processes*, 67 J. Personality & Soc. Psych. 818 (1994), https://doi.org/10.1037//0022-3514.67.5.818.

109. Elizabeth F. Loftus, Eyewitness Testimony (1979).

handgun at the expense of gathering precise information about the perpetrator's face.[110]

To get some sense of the prevalence and magnitude of the weapon focus effect in the eyewitness context, many applied studies have manipulated the presence of a weapon and measured eyewitness performance.[111] The consensus from well-designed studies is precisely what we should expect from the basic science of visual attention and inattentional blindness: Brandishing a weapon limits the ability of eyewitnesses to discriminate between perpetrators and innocent suspects,[112] and markedly impairs identification accuracy.[113]

Laboratory studies of weapon focus may significantly underestimate the magnitude of the effect in actual crimes, since the weapon has little threat value in a laboratory setting. Some efforts have been made to address this concern by correlating the presence of a weapon in actual crimes to identification accuracy,[114] but interpretation of these analyses is hampered by lack of experimental control over other variables and questionable assumptions of ground truth.[115]

Finally, because the weapon focus effect is an attention-based phenomenon, it bears on the utility of the *Manson* assertion that eyewitness reliability (the probability that the testimony is correct) can be predicted by "the witness' degree of attention." As reviewed above, basic research demonstrates that there are two components to attention: magnitude (or "degree") and directional focus. Magnitude of attention is indeed high under conditions in which a weapon is present, but the directional focus of attention in such cases precludes accurate identification of the perpetrator, because the eyewitness's attention is directed at the weapon and not the perpetrator's face.

Cross-Race Effect

The cross-race effect is a well-known phenomenon of visual cognition and perceptual learning, manifested as better performance on a task that involves

110. Loftus et al., *supra* note 62.

111. Nancy Mehrkens Steblay, *A Meta-Analytic Review of the Weapon Focus Effect*, 16 Law & Hum. Behav. 413 (1992), https://doi.org/10.1007/bf02352267; Jonathan M. Fawcett et al., *Of Guns and Geese: A Meta-Analytic Review of the 'Weapon Focus' Literature*, 19 Psych., Crim. & Law 35 (2011), https://doi.org/10.1080/1068316x.2011.599325.

112. Curt A. Carlson & Maria A. Carlson, *An Evaluation of Lineup Presentation, Weapon Presence, and a Distinctive Feature Using ROC Analysis*, 3 J. Applied Rsch. Memory & Cognition 45 (2014), https://doi.org/10.1016/j.jarmac.2014.03.004.

113. Curt A. Carlson & Maria A. Carlson, *A Distinctiveness-Driven Reversal of the Weapon-Focus Effect*, 6(1) Applied Psych. Crim. Just. 82 (2012), https://doi.org/10.1037/h0101806.

114. Fawcett et al., *supra* note 111, at 1–32.

115. *Identifying the Culprit, supra* note 12.

discriminating between faces of one's own race relative to faces of a different race.[116] The effect has been studied extensively in basic research on vision and memory (see sections titled "How Vision Works" and "How Memory Works" above) and is, in part, a by-product of categorical perception. The strength and context generality of this perceptual phenomenon offer good reasons to suspect that it will adversely impact the accuracy of eyewitness identification. Many laboratory studies of eyewitness identification over the past 50 years have confirmed this: The ability to discriminate between the culprit and an innocent suspect suffers and the accuracy of identifications is significantly impaired under conditions in which the eyewitness is of a different race from the lineup participants.[117] This impaired eyewitness recognition ability has also been demonstrated experimentally in noncriminal real-world contexts,[118] suggesting that it is not a laboratory artifact. Moreover, about half of all persons exonerated based on postconviction DNA testing were convicted, in part, based on mistaken cross-race identifications.[119] Applied eyewitness studies, together with extensive basic research on this topic, demonstrate that eyewitness testimony employed by police investigators and courts is, on average, less likely to be accurate if the witness and the suspect are of different races.

Viewing Distance, Lighting, and Exposure Duration

These three factors routinely affect the fidelity of visual experience, as revealed by basic research on the operating characteristics of human vision (see section titled "How Vision Works" above). In simple terms that comport with normal human experience, far viewing distances, dim lighting, and short viewing durations all limit the amount of visual information available for object recognition.

116. McGugin et al., *supra* note 15; Wong et al., *supra* note 15.

117. Christian A. Meissner & John C. Brigham, *Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review*, 7 Psych., Pub. Pol'y & L. 3 (2001), https://doi.org/10.1037//1076-8971.7.1.3; Jungwon Lee & Steven D. Penrod, *Three-Level Meta-Analysis of the Other-Race Bias in Facial Identification*, 36 Applied Cognitive Psych. 1106 (2022), https://doi.org/10.1002/acp.3997; Jacqueline Katzman & Margaret Bull Kovera, *Potential Causes of Racial Disparities in Wrongful Convictions Based on Mistaken Identifications: Own-Race Bias and Differences in Evidence-Based Suspicion*, 47 Law & Hum. Behav. 23 (2023), https://doi.org/10.1037/lhb0000503.supp; Rachel A. Wilson & Tammy L. Sonnentag, *The Effect of Race of the Perpetrator and Misinformation on Eyewitness Accuracy and Confidence*, 26 Mod. Psych. Studs. 8 (2021); Jesse N. Rothweiler, Kerri A. Goodwin & Jeff Kukucka, *Presence of Administrators Differentially Impacts Eyewitness Discriminability for Same–And Other–Race Identifications*, 34 Applied Cognitive Psych. 1530 (2020), https://doi.org/10.1002/acp.3733; Margaret Bull Kovera & Andrew J. Evelo, *Eyewitness Identification in its Social Context*, 10 J. Applied Rsch. Memory & Cognition 313 (2021), https://doi.org/10.1016/j.jarmac.2021.04.003.

118. John C. Brigham et al., *Accuracy of Eyewitness Identification in a Field Setting*, 42 J. Pers. & Soc. Psych. 673 (1982), https://doi.org/10.1037//0022-3514.42.4.673.

119. Garrett, *supra* note 2 at 73.

Unsurprisingly, they all affect eyewitness performance in predictable ways, as revealed by applied studies of eyewitness identification.

For example, one study found that it is harder to identify faces at a distance because of limits on spatial acuity.[120] Numerous other applied eyewitness studies similarly support the conclusion that as viewing distance increases, eyewitness performance worsens.[121] One study made the claim that eyewitness accuracy declines by approximately half when viewing distance increases from 3 to 20 meters.[122] The quantitative characteristics of this relationship (and those drawn simply from basic research on vision) suggest that reliable measures of viewing distance in a given case can help predict the accuracy of an identification.

There is also a large basic science literature on visibility of objects as a function of ambient illumination (see Figure 1), which naturally suggests that eyewitness performance should decline as lighting dims. Indeed it does.[123] Other studies have looked at the combined effects of distance and lighting, with one report offering a rule of thumb: Eyewitness performance is acceptable only if lighting exceeds 15 lux (roughly twilight) and distance is less than 15 meters.[124] Findings from these applied studies, in conjunction with known operating characteristics of human vision, can support quantitative estimates of the accuracy of eyewitness reports.

Finally, basic sciences have long established that duration of exposure (viewing time) to a visual stimulus is correlated with information gain.[125] Much of the impaired performance associated with short viewing durations is simply due to

120. Geoffrey R. Loftus & Erin M. Harley, *Why Is It Easier to Identify Someone Close than Far Away?*, 12 Psychonomic Bull. & Rev. 43 (2005), https://doi.org/10.3758/bf03196348.

121. Robert F. Lockamyeir et al., *The Effect of Viewing Distance on Empirical Discriminability and the Confidence–Accuracy Relationship for Eyewitness Identification*, 34 Applied Cognitive Psych. 1047 (2020), https://doi.org/10.1002/acp.3683; James Michael Lampinen et al., *Effects of Distance on Face Recognition: Implications for Eyewitness Identification*, 21 Psychonomic Bull. & Rev. 1489 (2014), https://doi.org/10.3758/s13423-014-0641-2; C.L. Lindsay et al., *How Variations in Distance Affect Eyewitness Reports and Identification Accuracy*, 32 Law & Hum. Behav. 526 (2008), https://doi.org/10.1007/s10979-008-9128-x; Thomas J. Nyman et al., *The Distance Threshold of Reliable Eyewitness Identification*, 43 Law & Hum. Behav. 527 (2019), https://doi.org/10.1037/lhb0000342.supp; Thomas J. Nyman et al., *A Stab in the Dark: The Distance Threshold of Target Identification in Low Light*, 6 Cogent Psych. 1632047 (2019), https://doi.org/10.1080/23311908.2019.1632047.

122. Lockamyeir et al., *supra* note 63.

123. Lisa DiNardo & David Rainey, *The Effects of Illumination Level and Exposure Time on Facial Recognition*, 41 Psych. Rec. 329 (1991), https://doi.org/10.1007/bf03395115; Marloes De Jong et al., *Familiar Face Recognition as a Function of Distance and Illumination: A Practical Tool for Use in the Courtroom*, 11 Psych., Crim. & L. 87 (2005), https://doi.org/10.1080/10683160410001715123.

124. Willem A. Wagenaar & Juliette H. Van Der Schrier, *Face recognition as a function of distance and illumination: A practical tool for use in the courtroom*, 2 Psych., Crim. & L. 321 (1996), https://doi.org/10.1080/10683169608409787.

125. George Sperling, *The Information Available in Brief Visual Presentations*, 74 Psych. Monographs: Gen. & Applied 1 (1960), https://doi.org/10.1037/h0093759.

loss of acuity and reduced contrast sensitivity.[126] Sufficiently long viewing times, by contrast, equip observers with the information needed to make better decisions about what they have seen.[127] Viewing durations are, of course, not under experimental control in actual eyewitness cases and, regrettably, they are not often recorded with precision,[128] but one report estimates that over a third of all eyewitness reports are based on viewing durations of less than one minute.[129] Numerous laboratory studies of eyewitness performance have manipulated viewing time and found, as expected, that longer viewing times are associated with greater eyewitness accuracy,[130] although high-confidence identifications are generally highly accurate for both short and long exposure durations,[131] an issue we take up in the section titled "Confidence" below.

126. Jacob Nachmias, *Effect of Exposure Duration on Visual Contrast Sensitivity with Square-Wave Gratings*, 57 J. Opt. Soc. Am. 421 (1967), https://doi.org/10.1364/josa.57.000421; J. Jason McAnany, *The Effect of Exposure Duration on Visual Acuity for Letter Optotypes and Gratings*, 105 Vision Rsch. 86 (2014), https://doi.org/10.1016/j.visres.2014.08.024.

127. Radoslaw Martin Cichy, Dimitrios Pantazis & Aude Oliva, *Resolving Human Object Recognition in Space and Time*, 17 Nature Neurosci. 455 (2014), https://doi.org/10.1038/nn.3635.

128. An eyewitness may be the only source of information regarding the duration of exposure to the perpetrator during a crime. Evidence suggests that these estimates are rarely accurate. Holly L. Gasper, Michael M. Roy & Heather D. Flowe, *Improving Time Estimation in Witness Memory*, 10 Frontiers Psych. 1452 (2019), https://doi.org/10.3389/fpsyg.2019.01452; A. Daniel Yarmey & Meagan J. Yarmey, *Eyewitness Recall and Duration Estimates in Field Settings*, 27 J. Applied Soc. Psych. 330 (1997), https://doi.org/10.1111/j.1559-1816.1997.tb00635.x.

129. Tim Valentine, Alan Pickering & Stephen Darling, *Characteristics of Eyewitness Identification That Predict the Outcome of Real Lineups*, 17 Applied Cognitive Psych. 969 (2003), https://doi.org/10.1002/acp.939.

130. J. Kirkland Reynolds & Kathy Pezdek, *Face Recognition Memory: The Effects of Exposure Duration and Encoding Instruction*, 6 Applied Cognitive Psych. 279 (1992), https://doi.org/10.1002/acp.2350060402; Amina Memon, Lorraine Hope & Ray Bull, *Exposure Duration: Effects on Eyewitness Accuracy and Confidence*, 94 Brit. J. Psych. 339 (2003), https://doi.org/10.1348/000712603767876262; Matthew A. Palmer et al., *The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention*, 19 J. Exper. Psych.: Applied 55 (2013), https://doi.org/10.1037/a0031602; Rolando N. Carol & Nadja Schreiber Compo, *The Effect of Encoding Duration on Implicit and Explicit Eyewitness Memory*, 61 Consciousness & Cognition 117 (2018), https://doi.org/10.1016/j.concog.2018.02.004; Brian H. Bornstein et al., *Effects of Exposure Time and Cognitive Operations on Facial Identification Accuracy: A Meta-Analysis of Two Variables Associated with Initial Memory Strength*, 18 Psych., Crim. & L. 473 (2012), https://doi.org/10.1080/1068316x.2010.508458; Neil Brewer, Michael Gordon & Nigel Bond, *Effect of Photoarray Exposure Duration on Eyewitness Identification Accuracy and Processing Strategy*, 6 Psych., Crim. & L. 21 (2000), https://doi.org/10.1080/10683160008410829.

131. Carolyn Semmler et al., *The Role of Estimator Variables in Eyewitness Identification*, 24 J. Exper. Psych.: Applied 400 (2018), https://doi.org/10.1037/xap0000157; Matthew A. Palmer et al., *The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention*, 19 J. Exper. Psych.: Applied 55 (2013), https://doi.org/10.1037/a0031602.supp.

The voluminous body of data on exposure duration thus demonstrates that, when reliable information about viewing times is available, this information can be used to infer the probability that eyewitness testimony is accurate.

Multiple Perpetrators

Crimes are often carried out by more than one person. Evidence indicates that the number of multiperpetrator homicides, for example, has been on the rise for decades and is estimated to exceed 20% of all murders.[132] Explanations point to growing gang violence, peer pressure, and the perception of distributed responsibility ("de-individuation" of any single perpetrator)—a form of groupthink that licenses behaviors that individuals acting alone would view as impermissible.

For our purposes, the important empirical question is: How well can a witness be expected to remember one or more faces under these circumstances? This is a divided attention problem, much like the weapon focus problem addressed above, in which limited attentional resources must be distributed across multiple targets. While focused attention generally improves the fidelity of vision and memory for an attended stimulus, considerable evidence from basic science demonstrates what comports with everyday human experience: Detection, discriminability, and recognition of visual targets is impaired when attention must be subdivided between them.[133] A number of laboratory studies have explored the consequences of this phenomenon for eyewitness identification.[134] As expected (and mirroring the weapon focus effect on divided attention), identification accuracy is consistently poorer when multiple perpetrators are

132. Lockamyeir et al., *supra* note 63; Nortje et al., *supra* note 63, at 348–82.

133. Edward Awh & Harold Pashler, *Evidence for Split Attentional Foci*, 26 J. Exper. Psych. Hum. Perception & Performance 834 (2000), https://doi.org/10.1037//0096-1523.26.2.834; Arthur F. Kramer & Sowon Hahn, *Splitting the Beam: Distribution of Attention over Noncontiguous Regions of the Visual Field*, 6 Psych. Sci. 381 (1995), https://doi.org/10.1111/j.1467-9280.1995.tb00530.x; Won Mok Shim, George A. Alvarez & Yuhong V. Jiang, *Spatial Separation Between Targets Constrains Maintenance of Attention on Multiple Objects*, 15 Psychonomic Bull. & Rev. 390 (2008), https://doi.org/10.3758/pbr.15.2.390; Steven Yantis, *Multi-Element Visual Tracking: Attention and Perceptual Organization*, 24 Cognitive Psych. 295 (1992), https://doi.org/10.1016/0010-0285(92)90010-y; Amelia H. Harrison, Sam Ling & Joshua J. Foster, *The Cost of Divided Attention for Detection of Simple Visual Features Primarily Reflects Limits in Post-Perceptual Processing*, 85 Attention, Perception, & Psychophysics 377 (2023), https://doi.org/10.3758/s13414-022-02547-7; Harold Pashler, *Attention and Visual Perception: Analyzing Divided Attention*, *in* Visual Cognition: An Invitation to Cognitive Science 71 (2d ed. 1995), https://doi.org/10.7551/mitpress/3965.003.0005; M. Corbetta et al., *Selective and Divided Attention During Visual Discriminations of Shape, Color, and Speed: Functional Anatomy by Positron Emission Tomography*, 11 J. Neurosci. 2383 (1991), https://doi.org/10.1523/jneurosci.11-08-02383.1991; Nilli Lavie, *Distracted and Confused? Selective Attention Under Load*, 9 Trends Cognitive Sci. 75 (2005), https://doi.org/10.1016/j.tics.2004.12.004.

134. Clifford & Hollin, *supra* note 63; Hobson & Wilcock, *supra* note 63; Lockamyeir et al., *supra* note 63.

witnessed during a crime. For the purposes of the courts, it follows that eyewitness testimony under these conditions is less likely to be correct.

Facial Recognition Ability

Basic research shows (see section titled "Categorical perception" above) that one feature of categorical visual perception is category-specific expertise. While humans have a highly specialized system for face processing,[135] face recognition ability varies significantly across the population, from "super recognizers" at one extreme to "face-blind people" at the other (see Figure 12).[136] Chad Dodson and colleagues hypothesized that knowledge of the facial recognition ability of a witness might provide insight into the probative value of an identification by that witness.[137] Results from this laboratory study reveal that a high-confidence identification is far more likely to be in error if the witness scores poorly on a test of facial recognition ability. These findings indicate that a simple test of face recognition ability can provide factfinders with valuable insight into the accuracy of high confidence identifications. We would not hesitate to conduct a visual acuity test to explore whether poor eyesight could have affected an eyewitness identification. Similarly, testing the face recognition ability of an eyewitness could become a standard procedure for law enforcement and for courts, which serves the purpose of inferring identification accuracy.

Retention Interval

As noted above (see section titled "Memory Storage"), many decades of basic research on memory have revealed that stored information is lost with the passage of time, as evidenced by progressive performance deterioration on standard visual discrimination and recognition tasks (see Figure 11). This memory loss is most rapid immediately after acquisition, when consolidation into long-term memory is not yet complete. After this initial period, consolidation continues but resulting long-term memories are readily modified by exposure to new information, the likelihood of which naturally increases with time. A witness's inevitable interactions with law enforcement and legal counsel, not to mention communications from journalists, family, and friends, have the potential to significantly modify the witness's memory of faces encountered and other details of

135. Winrich Freiwald, Bradley Duchaine & Galit Yovel, *Face Processing Systems: From Neurons to Real World Social Perception*, 39 Ann. Rev. Neurosci. 325 (2016), https://doi.org/10.1146/annurev-neuro-070815-013934.

136. Duchaine & Nakayama, *supra* note 101.

137. Grabman et al., *supra* note 103, at 233–43.

the crime scene.[138] Thus, the fidelity of retrieved events—and the accuracy of identification—is likely to be greater when retrieval occurs closer to the time of the witnessed events.

These concerns about the completeness and stability of long-term memories have motivated a large number of applied studies of performance on eyewitness identification as a function of retention interval, defined as the length of time from the witnessed event to the lineup. A meta-analysis revealed that performance declines significantly as retention interval increases.[139] Moreover, the authors maintain that the rate of memory loss, modeled as a power law,[140] can be used to estimate the probability that an identification is correct: "[N]ot only can the percentage of remaining memory strength be determined for any retention interval, but this strength estimate can be translated into an estimated probability of being correct on a fair lineup of a specified size."[141]

The initial strength of an encoded face memory is, of course, modulated by other estimator variables associated with witnessing the crime, such as cognitive state (e.g., attention and emotion), viewing conditions (e.g., distance, lighting, and exposure duration), and the cross-race effect.[142] Nonetheless, there are strategies for estimating initial memory strength, and one study makes the promising claim that "we can now offer the judge or juror an estimate of what proportion of memory strength, regardless of its initial value, remained at the time the eyewitness's memory for the perpetrator was tested."[143] This predictive approach could provide a quantitative foundation for the fifth *Manson* criterion for accuracy prediction: "the length of time between the crime and the confrontation."

Stress and Emotion

As noted above in our review of the basic sciences of vision and memory (see sections titled "Visual attention" and "The effect of emotion on memory storage" above), considerable empirical evidence demonstrates that stress and high states of

138. Maria S. Zaragoza & Sean M. Lane, *Source Misattributions and the Suggestibility of Eyewitness Memory*, 20 J. Exper. Psych.: Learning, Memory & Cognition 934 (1994), https://doi.org/10.1037//0278-7393.20.4.934; William C. Thompson, et al., *What Did the Janitor Do? Suggestive Interviewing and the Accuracy of Children's Accounts*, 21 Law & Hum. Behav. 405 (1997), https://doi.org/10.1023/a:1024859219764; D. Stephen Lindsay & Marcia K. Johnson, *The Eyewitness Suggestibility Effect and Memory for Source*, 17 Memory & Cognition 349 (1989), https://doi.org/10.3758/bf03198473.

139. Deffenbacher et al. *supra* note 26, at 139.

140. Wickelgren, *supra* note 85; Wixted & Ebbesen, *supra* note 85.

141. Deffenbacher et al., *supra* note 26.

142. Melissa A. Pigott, John C. Brigham & Robert K. Bothwell, *A Field Study on the Relationship Between Quality of Eyewitnesses' Descriptions and Identification Accuracy*, 17 J. Police Sci. & Admin. 84 (1990), https://perma.cc/9GXP-8DWH.

143. Deffenbacher et al., *supra* note 26.

emotion can influence the allocation of visual attention as well as the encoding and retrieval of memories. Witnessing a crime, and the subsequent identification event, frequently elicit stress and often trigger high levels of emotion (commonly fear)—particularly in cases where the witness is a victim, a weapon is involved, or there is serious threat to life or limb. It naturally follows that these altered states of the observer may affect the content and validity of eyewitness reports.

The effects of stress and emotion on eyewitness performance have been difficult to study, as levels that often accompany eyewitness events in the real world are extremely difficult to replicate in a laboratory setting, in part because of the obvious pretense and the fact that regulations prohibit induction of high stress levels in human subjects. Some studies have, nonetheless, reportedly achieved moderate levels of stress and fear using unusual techniques, which generally involve manipulations that are ancillary to the identification test itself. One experiment involved subjecting active-duty military personnel to a mock POW camp (a valid part of military survival school training), which included aggressive interrogation and physical isolation-related stress.[144] The study found that memories acquired during stressful events are highly vulnerable to modification by exposure to postevent misinformation, even in individuals whose level of training and experience might be considered relatively immune to such influences. Other studies of this sort have attempted to elicit stress and fear by placing subjects serving as witnesses in a creepy environment (the Horror Labyrinth of the London Dungeon)[145] or administering vaccine injections to children who were serving as witnesses.[146]

Another approach in laboratory studies has been to independently confirm that "putative manipulations of stress/anxiety or perceived level of violence had been successful."[147] One meta-analysis limited to studies that obtained such confirmation (using standard physiological measures, such as breathing, and heart rate) found, as expected, that confirmed states of stress were correlated with impairments of identification accuracy and impaired ability to identify key characteristics of a witnessed individual's face (e.g., hair length, hair color, eye color, shape of face, presence of facial hair). As emphasized in a more recent review, "ensuring valid stress inductions is a prerequisite for effectively studying effects of acute stress on memory performance."[148]

144. C.A. Morgan III et al., *Misinformation Can Influence Memory for Recently Experienced Highly Stressful Events*, 36 Int'l J.L. & Psychiatry 11 (2013), https://doi.org/10.1016/j.ijlp.2012.11.002.

145. Tim Valentine & Jan Mesout, *Eyewitness Identification Under Stress in the London Dungeon*, 23 Applied Cognitive Psych. 151 (2009), https://doi.org/10.1002/acp.1510.

146. Douglas P. Peters, *Stress, Arousal, and Children's Eyewitness Memory*, *in* Memory for Everyday and Emotional Events 351 (1st ed. 2018), https://doi.org/10.4324/9781315789231-15.

147. Kenneth A. Deffenbacher et al., *A Meta-Analytic Review of the Effects of High Stress on Eyewitness Memory*, 28 Law & Hum. Behav. 687 (2004), https://doi.org/10.1007/s10979-004-0565-x.

148. Carey Marr et al., *The Effects of Acute Stress on Eyewitness Memory: An Integrative Review for Eyewitness Researchers*, 29 Memory 1091 (2021), https://doi.org/10.1080/09658211.2021.1955935.

Laboratory studies have also revealed a counterintuitive relationship between accuracy and confidence under conditions of stress. Specifically, while stress clearly interferes with the accuracy of eyewitness memory, the predictive relationship between confidence and accuracy is similar across stressful and stress-free viewing conditions. (Stress-free observers are highly confident when accurate, while stressed observers are not confident when inaccurate.) This stability of the confidence–accuracy relationship implies that stressed witnesses are sufficiently aware of their cognitive state, and its potential to limit accuracy, and they "scale back their confidence judgments" accordingly.[149] These findings imply that confidence reports may be useful predictors of accuracy under conditions of stress.

Despite the promise of these various approaches, it is impossible under laboratory conditions to achieve levels of stress and fear experienced by, for example, real witness–victims of rape, or a witness to the murder of a relative or friend, or a mass shooting. Nonetheless, although *Manson* made no mention of stress or fear in their predictive framework, the existing data argue that if stress or fear can be inferred from the circumstances of a witnessed crime, any eyewitness testimony proffered to the court will be of questionable accuracy.

Confidence

No topic in the literature on eyewitness identification has elicited as much robust discussion and research as confidence and its relationship to accuracy.[150] In broad terms, this debate stems from the conflicting goals of (a) finding a measure that can predict the accuracy of an identification, and (b) avoiding the enormous personal and societal risk that such a measure might fail.

Before we review the basis for this debate, it is useful to define what we mean by confidence in this context and illustrate how it works. Confidence has a broad colloquial definition, but in the eyewitness context it refers to the perceived degree of certainty in a piece of information or a decision. Confidence normally grows when multiple sources of information paint a coherent picture.[151] Many decades of basic research on confidence and the accuracy of

149. Sara D. Davis et al., *Physiological Stress and Face Recognition: Differential Effects of Stress on Accuracy and the Confidence–Accuracy Relationship*, 8 J. Applied Rsch. Memory & Cognition 367 (2019), https://doi.org/10.1016/j.jarmac.2019.05.006.

150. Confidence has traditionally been treated as an estimator variable in applied eyewitness research, since it is a property of the witness under the conditions of viewing. As this science has advanced, however, it has become clear that confidence as a variable is not so easily pigeonholed: It can be predictably influenced by the state of systems variables.

151. Kahneman & Tversky, *supra* note 77, at 237–51; Daniel Kahneman, Thinking, Fast and Slow (2011).

retrieved memories reveal that the correlation is frequently poor,[152] which we can largely attribute to the natural process of smoothing over source discrepancies to create a sense of coherence, which, in turn, fosters confidence. This dissociation between confidence and accuracy has obvious implications for eyewitness testimony, and many have warned of the risks associated with the use of witness confidence statements for criminal investigation and prosecution.[153]

Timing of identification and conditions of confidence assessment

Applied eyewitness research has revealed nuances of the confidence–accuracy relationship. For example, the problem of confidence inflation,[154] in which witnesses progressively gain greater confidence during the window of time between the initial lineup identification and the courtroom identification, has led to the recommendation that only the initial identification and confidence statements made at that time be used as evidence.[155] Considerable evidence supports this

152. Loftus, *supra* note 109; Loftus & Pickrell, *supra* note 88, at 720–25; Kenneth A. Norman & Daniel L. Schacter, *False Recognition in Younger and Older Adults: Exploring the Characteristics of Illusory Memories*, 25 Memory & Cognition 838 (1997), https://doi.org/10.3758/bf03211328; Henry L. Roediger III & Kathleen B. McDermott, *Creating False Memories: Remembering Words Not Presented in Lists*, 21 J. Exper. Psych.: Learning, Memory & Cognition 803 (1995), https://doi.org/10.1037//0278-7393.21.4.803; Thomas A. Busey et al., *Accounts of the Confidence–Accuracy Relation in Recognition Memory*, 7 Psychonomic Bull. & Rev. 26 (2000) https://doi.org/10.3758/BF03210724; Daniel L. Schacter & Chad S. Dodson, *Misattribution, False Recognition and the Sins of Memory*, 356 Phil. Transactions Royal Soc. London Series B: Biological Scis. 1385 (2001), https://doi.org/10.1098/rstb.2001.0938; Hal R. Arkes, *Overconfidence in Judgmental Forecasting*, *in* Principles of Forecasting 495 (2001), https://doi.org/10.1007/978-0-306-47630-3_22; Mark Tippens Reinitz et al., *Different Confidence–Accuracy Relationships for Feature-Based and Familiarity-Based Memories*, 37 J. Exper. Psych.: Learning, Memory & Cognition 507 (2011), https://doi.org/10.1037/a0021961.supp; Henry L. Roediger III, John H. Wixted & K. Andrew DeSoto, *The Curious Complexity between Confidence and Accuracy in Reports from Memory*, in Memory and Law 84–118 (Lynn Nadel & Walter Sinnott-Armstrong eds., 2012); Oxford University Press, New York; Elizabeth F. Loftus & Rachel L. Greenspan, *If I'm Certain, Is It True? Accuracy and Confidence in Eyewitness Memory*, 18 Psych. Sci. Pub. Int. 1 (2017), https://doi.org/10.1177/1529100617699241.

153. Sporer et al., *supra* note 75; Deffenbacher, *supra* note 75; Reinitz et al., *supra* note 75; Bothwell et al., *supra* note 75.

154. Melissa Paiva et al., *Influence of Confidence Inflation and Explanations for Changes in Confidence on Evaluations of Eyewitness Identification Accuracy*, 16 Legal & Criminological Psych. 266 (2011), https://doi.org/10.1348/135532510x503340; Sporer et al., *supra* note 75; Deffenbacher, *supra* note 75; Reinitz et al., *supra* note 75; Bothwell et al., *supra* note 75; Laura Smalarz & Gary L. Wells, *Do Multiple Doses of Feedback Have Cumulative Effects on Eyewitness Confidence?*, 9 J. Applied Rsch., Memory & Cognition 508 (2020), https://doi.org/10.1037/h0101857; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications*, 20 Psych., Pub. Pol'y & L. 1 (2014), https://doi.org/10.1037/law0000001; Douglass & Steblay, *supra* note 75.

155. Identifying the Culprit, *supra* note 12.

recommendation: Memory is patently malleable,[156] and the very act of probing memory in a lineup procedure can actually modify that memory in a way that will lessen the accuracy of future identifications.[157] Following this logic, an argument can be made that the initial identification should be the *only* identification obtained.[158]

Additional insight into this issue has come from work by John Wixted and colleagues, who argued that confidence reports made at the time of the initial identification have their greatest predictive value if they are held to a very high standard. Specifically, these investigators report that the relationship between confidence and accuracy is strong if the initial identification is made under "pristine" conditions.[159] Those conditions include a properly blinded test and fair lineup construction (no participant stands out), which promote coherence of information received by the eyewitness, thus limiting the accuracy-reducing inclination to smooth over discrepancies.

Critics argue that it may be difficult to recognize pristine conditions in real casework, owing in part to limited access to situational context (e.g., what exactly did the police do?). By this argument, eyewitness confidence may place innocent suspects at great risk.[160] Some disagreement remains on this point,[161] but the rationale for the pristine distinction and the use of only the initial identification and confidence statement rests firmly on the cognitive science of memory and decision-making.[162]

156. Deborah Davis & Elizabeth F. Loftus, *Eyewitness Science in the 21st Century: What Do We Know and Where Do We Go from Here?*, *in* 1 Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th ed. 2018), https://doi.org/10.1002/9781119170174.epcn116.

157. Nancy K. Steblay, Robert W. Tix & Samantha L. Benson, *Double Exposure: The Effects of Repeated Identification Lineups on Eyewitness Accuracy*, 27 Applied Cognitive Psych. 644 (2013), https://doi.org/10.1002/acp.2944; Nancy K. Steblay & Jennifer E. Dysart, *Repeated Eyewitness Identification Procedures with the Same Suspect*, 5 J. Applied Rsch. Memory & Cognition 284 (2016), https://doi.org/10.1016/j.jarmac.2016.06.010.

158. John T. Wixted et al., *Test a Witness's Memory of a Suspect Only Once*, 22 Psych. Sci. Pub. Int. 1S (2021), https://doi.org/10.1177/15291006211026259.

159. *Id.*; Wixted & Wells, *supra* note 25; Semmler et al., *supra* note 131.

160. Shari R. Berkowitz & Steven J. Frenda, *Rethinking the Confident Eyewitness: A Reply to Wixted, Mickes, and Fisher*, 13 Persps. Psych. Sci. 336 (2018), https://doi.org/10.1177/1745691617751883; Shari R. Berkowitz et al., *Convicting with Confidence? Why We Should Not Over-Rely on Eyewitness Confidence*, 30 Memory 10 (2022), https://doi.org/10.1080/09658211.2020.1849308; Shari R. Berkowitz et al., *Eyewitness Confidence May Not be Ready for the Courts: A Reply to Wixted Et Al.*, 30 Memory 75 (2022), https://doi.org/10.1080/09658211.2021.1952271; James D. Sauer, Matthew A. Palmer & Neil Brewer, *Pitfalls in Using Eyewitness Confidence to Diagnose the Accuracy of an Individual Identification Decision*, 25 Psych., Pub. Pol'y & L. 147 (2019), https://doi.org/10.1037/law0000203.

161. Travis M. Seale-Carlisle et al., *Confidence and Response Time as Indicators of Eyewitness Identification Accuracy in the Lab and in the Real World*, 84 J. Applied Rsch. Memory & Cognition 420 (2019), https://doi.org/10.1016/j.jarmac.2019.09.003; Wixted & Wells, *supra* note 25.

162. Semmler et al., *supra* note 131.

System Variables

System variables are causal states that are potentially under the control of the criminal justice system, since they bear on activities that occur after the witnessed events. These include (1) communication between law enforcement and the witness, (2) communication between the witness and the larger community (legal defense, media, family and friends, etc.), (3) the practice of "evidence-based suspicion," (4) prior exposure to images of potential suspects, (5) the type of procedure used for identification, and (6) the choice of lineup fillers (lineup participants known to be innocent).

The empirical findings and recommendations derived from research on system variables are of direct relevance to, and advisory of, police practices. That is, they can be used to improve eyewitness identification performance. Knowledge of the states of these variables for a given case, however, can also be useful to the courts in estimating the probability that eyewitness testimony gained by law enforcement is accurate.

Communication Between Law Enforcement Personnel and Witness

Statements that convey prior probability of suspicion

Police commonly have information that could influence both the accuracy of a witness's identification and the confidence expressed in the identification. For example, a police officer might first report to the witness/victim of a carjacking that a suspect was apprehended in the stolen vehicle and then ask the witness to view a lineup with that suspect in it. This might seem innocuous on the surface and could be viewed as reinforcing information for the witness, but it establishes a significant prior probability that the perpetrator is in the lineup. Inferences drawn from such priors can readily bias perception, choice, and action under conditions of uncertainty.[163] The result is that the witness may actually *perceive* the perpetrator in a target-absent lineup and may unwittingly adopt a lower criterion for reporting an identification even when recognition memory is poor.

Communication during lineups: The importance of blinded lineup administration

In cases where the administrator of a lineup knows the status of lineup participants, they are in a position to inadvertently communicate that information to a

163. Albright, *supra* note 36.

witness. The witness, commonly eager to be of assistance, can easily pick up unintended signals (e.g., direction of gaze, body posture, facial expressions), which may be highly suggestive and have the effect of biasing the outcome—yielding a larger fraction of misidentifications and inflation of witness confidence. Generally, the solution to these biasing effects is to limit communication between the witness and people with relevant knowledge.[164]

Many applied eyewitness studies support this information-control approach and demonstrate that the problem of inadvertent communication during a lineup can be eliminated by the practice of "double blinding."[165] This simple and effective tool is a pillar of the scientific method.[166] In an experimental test of a novel drug, for example, a subject is prevented (hence "blinded" or "single blinded") from knowing the experimental condition that they have been assigned to. Furthermore, the scientist who measures the effect of the drug is also prevented ("double blinded") from knowing the experimental condition associated with that measure, all of which precludes bias based on prior knowledge. Similarly, in a lineup, the witness is naturally blinded to knowledge of the "conditions," but the risk of bias can be reduced if the lineup administrator has no knowledge to convey—that is, the test is double blinded. This simple practice is highly effective at improving eyewitness accuracy, and the risks of not conducting double-blinded lineups are clear. For all of these reasons, failure to do so runs afoul of

164. Ryann M. Haw & Ronald P. Fisher, *Effects of Administrator-Witness Contact on Eyewitness Identification Accuracy*, 89 J. Applied Psych. 1106 (2004), https://doi.org/10.1037/0021-9010.89.6.1106; Robert Rosenthal, *Covert Communication in Classrooms, Clinics, Courtrooms, and Cubicles*, 57 Am. Psych. 839 (2002), https://doi.org/10.1037//0003-066x.57.11.839.

165. Steven E. Clark, Tanya E. Marshall & Robert Rosenthal, *Lineup Administrator Influences on Eyewitness Identification Decisions*, 15 J. Exper. Psych.: Applied 63 (2009), https://doi.org/10.1037/a0015185; Margaret Bull Kovera & Andrew J. Evelo, *The Case for Double-Blind Lineup Administration*, 23 Psych., Pub. Pol'y & L. 421 (2017), https://doi.org/10.1037/law0000139; Margaret Bull Kovera & Andrew J. Evelo, *Improving Eyewitness-Identification Evidence Through Double-Blind Lineup Administration*, 29 Current Directions Psych. Sci. 563 (2020), https://doi.org/10.1177/0963721420969366; Steve D. Charman & Vanessa Quiroz, *Blind Sequential Lineup Administration Reduces Both False Identifications and Confidence in Those Identifications*, 40 Law & Hum. Behav. 477 (2016), https://doi.org/10.1037/lhb0000197; Dario N. Rodriguez & Melissa A. Berry, *Eyewitness Science and the Call for Double-Blind Lineup Administration*, 2013 J. Criminology 530523 (2013), https://doi.org/10.1155/2013/530523; Jacqueline L. Austin et al., *Double-Blind Lineup Administration: Effects of Administrator Knowledge on Eyewitness Decisions*, *in* 139 Reform of Eyewitness Identification Procedures (2013), https://doi.org/10.1037/14094-007; Arthur H. Perlini & Andrew D. Silvaggio, *Eyewitness Misidentification: Single vs. Double-Blind Comparison of Photospread Administration*, 100 Psych. Reps. 247 (2007), https://doi.org/10.2466/pr0.100.1.247-256.

166. Michael Stolberg, *Inventing the Randomized Double-Blind Trial: The Nuremberg Salt Test of 1835*, 99 J. Royal Soc. Med. 642 (2006), https://doi.org/10.1177/014107680609901216; Lorraine Daston, *Scientific Error and the Ethos of Belief*, 72 Soc. Rsch.: Int'l Q. 1 (2005), https://doi.org/10.1353/sor.2005.0016; Paul J. Karanicolas, Forough Farrokhyar & Mohit Bhandari, *Practical Tips for Surgical Research: Blinding: Who, What, When, Why, How?*, 53 Can. J. Surg. 345 (2010); Gary L. Wells & C.A. Elizabeth Luus, *Police Lineups as Experiments: Social Methodology as a Framework for Properly Conducted Lineups*, 16 Pers. & Soc. Psych. Bull. 106 (1990), https://doi.org/10.1177/0146167290161008.

best practices,[167] recommendations of the U.S. Department of Justice,[168] and the International Association of Chiefs of Police.[169] In a court of law, lineup blinding status is naturally a useful predictor of eyewitness accuracy.

Video recording of lineup procedures

In view of the risk that some forms of communication between the lineup administrator(s) and witness can bias the identification, many have called for video recording of the entire procedure.[170] One may not expect that doing so will change the behavior of the people involved, but the recording serves as stable evidence to confirm—to the court, to the prosecution, and to the defense—whether the procedure was done in such a way to avoid a biased outcome. It also documents who was present, what the witness actually said about an identification and their level of confidence, and other behavioral characteristics (e.g., time to decide) that may bear on the accuracy of identifications. The absence of such video recordings, conversely, limits inferences about the accuracy of eyewitness testimony.

Postidentification feedback

Postidentification feedback from law enforcement can also influence a witness's confidence in their decision.[171] Applied studies have shown that simple reinforcing statements, even without specifics—"Nice work!"—can increase or decrease witness confidence, thus distorting any relationship to accuracy (see section titled "Timing of identification and conditions of confidence assessment" above).[172]

167. Identifying the Culprit, *supra* note 12.

168. U.S. Dep't of Just., Eyewitness Identification, *supra* note 7.

169. IACP (International Association of Chiefs of Police) Law Enforcement Policy Center, Eyewitness Identification: Model Policy Concepts & Issues (2016), https://perma.cc/3YDT-8A77.

170. Identifying the Culprit, *supra* note 12; Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3–36. U.S. Dep't of Just., Eyewitness Identification, *supra* note 7.

171. Douglass & Steblay, *supra* note 75; Mitchell L. Eisen et al., *Does Anyone Else Look Familiar? Influencing Identification Decisions by Asking Witnesses to Re-Examine the Lineup*, 42 Law & Hum. Behav. 306 (2018), https://doi.org/10.1037/lhb0000291.supp.

172. Carolyn Semmler, Neil Brewer & Gary L. Wells, *Effects of Postidentification Feedback on Eyewitness Identification and Nonidentification Confidence*, 89 J. Applied Psych. 334 (2004), https://doi.org/10.1037/0021-9010.89.2.334; Gary L. Wells & Amy L. Bradfield, *"Good, You Identified the Suspect": Feedback to Eyewitnesses Distorts Their Reports of the Witnessing Experience*, 83 J. Applied Psych. 360 (1998), https://doi.org/10.1037//0021-9010.83.3.360; Amy L. Bradfield, Gary L. Wells & Elizabeth A. Olson, *The Damaging Effect of Confirming Feedback on the Relation Between Eyewitness Certainty and Identification Accuracy*, 87 J. Applied Psych. 112 (2002), https://doi.org/10.1037//0021-9010.87.1.112; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post*

Very high or very low confidence expressed by a witness can, in turn, be very compelling to the trier of fact, regardless of whether the identification is correct.[173]

Communication Between Community and Witness: Social Influences on Eyewitness Performance

As we have seen, memory is malleable and the criteria that people use for momentous decisions can change over time. People are highly social creatures—eager for social reinforcement and easily swayed by pressure for social conformity—and it should come as no surprise that in the eyewitness context (and more generally), the expectations, real or inferred, of other parties (police, attorneys, news media, family, colleagues and comrades, social media followers) can have a profound influence over what we remember, the decisions we make, and the confidence we express. There are two important features to this potential for bias: the social utility of misinformation and the scale of social networks.

The social utility of misinformation

Social influences can readily modify an individual's memory of events. New information received from one's social networks may be unvetted "misinformation," which is seen as widely accepted by others and thus not held to rigorous standards. People have a regrettable affinity for such information,[174] in part because knowing things, regardless of whether they are true, has social value. The end result is contamination of the original content of memory, such that decisions are less accurate and confidence in them becomes distorted. In the

Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications, 20 Psych., Pub. Pol'y & L. 1 (2014), https://doi.org/10.1037/law0000001.

173. Steven D. Penrod & Brian L. Cutler, *Eyewitness Expert Testimony and Jury Decisionmaking*, 52 Law & Contemp. Probs. 43 (1989), https://doi.org/10.2307/1191907; Neil Brewer & Anne Burke, *Effects of Testimonial Inconsistencies and Eyewitness Confidence on Mock-Juror Judgments*, 26 Law & Hum. Behav. 353 (2002), https://doi.org/10.1023/a:1015380522722; James D. Sauer, Matthew A. Palmer & Neil Brewer, *Mock-Juror Evaluations of Traditional and Ratings-Based Eyewitness Identification Evidence*, 41 Law & Hum. Behav. 375 (2017), https://doi.org/10.1037/lhb0000235.supp; Crystal R. Slane & Chad S. Dodson, *Eyewitness Confidence and Mock Juror Decisions of Guilt: A Meta-Analytic Review*, 46 Law & Hum. Behav. 45 (2022), https://doi.org/10.1037/lhb0000481.supp.

174. Ilan Yaniv & Dean P. Foster, *Graininess of Judgment Under Uncertainty: An Accuracy-Informativeness Trade-Off*, 124 J. Exper. Psych.: Gen. 424 (1995), https://doi.org/10.1037/0096-3445.124.4.424; Rakefet Ackerman & Morris Goldsmith, *Control over Grain Size in Memory Reporting—With and Without Satisficing Knowledge*, 34 J. Exper. Psych.: Learning, Memory & Cognition 1224 (2008), https://doi.org/10.1037/a0012938.

extreme, people can hold and act upon "false memories" of things that never happened.[175]

Applied studies have explored the influence of postwitnessing information gain on eyewitness accuracy and confidence. The most straightforward of these effects is incorporation of new information into the preexisting memory of the crime scene, without awareness (a form of source memory failure).[176] Acquired knowledge of the eyewitness reports of others, for example, can bring about both a change in a witness's memory and confidence.[177] Perhaps more pernicious are influences from news reports, social media, and colleagues: Eyewitnesses in laboratory studies have been shown to readily modify their own memories, without awareness, to reflect information contained in news reports or implied by interviewer questions.[178]

A related line of research has focused on the effect of postwitnessing information on the inclination of a witness to be accurate versus informative.[179] Witnesses who are exposed to media reports that either promote the fidelity of eyewitness reports or denigrate it adopt a practice that comports.[180] Those who learn that eyewitness testimony is beneficial to criminal justice are inclined to be highly informative with a lower criterion for accuracy. By contrast, those who learn that eyewitness testimony is of questionable value are more circumspect and inclined toward accuracy.

175. Loftus, *supra* note 109; Loftus, *supra* note 88.

176. Lindsay & Johnson, *supra* note 138, at 349–58.

177. C. A. Elizabeth Luus & Gary L. Wells, *The Malleability of Eyewitness Confidence: Co-Witness and Perseverance Effects*, 79 J. Applied Psych. 714 (1994), https://doi.org/10.1037//0021-9010.79.5.714; Mitchell L. Eisen et al., *"I Think He Had a Tattoo on His Neck": How Co-Witness Discussions About a Perpetrator's Description Can Affect Eyewitness Identification Decisions*, 6 J. Applied Rsch. Memory & Cognition 274 (2017), https://doi.org/10.1016/j.jarmac.2017.01.009; Anders Granhag Pär et al., *Social Influence on Eyewitness Memory*, *in* Forensic Psychology in Context: Nordic and International Approaches 139 (2010), https://doi.org/10.4324/9781315094038-8.

178. Elizabeth F. Loftus & Guido Zanni, *Eyewitness Testimony: The Influence of the Wording of a Question*, 5 Bull. Psychonomic Soc. 86 (1975), https://doi.org/10.3758/bf03336715; Elizabeth F. Loftus, *Leading Questions and the Eyewitness Report*, 7 Cognitive Psych. 560 (1975), https://doi.org/10.1016/0010-0285(75)90023-7; Elizabeth F. Loftus, David G. Miller & Helen J. Burns, *Semantic Integration of Verbal Information into a Visual Memory*, 4 J. Exper. Psych.: Hum. Learning & Memory 19 (1978), https://doi.org/10.1037/0278-7393.4.1.19; Robin Blom & Kuo-Ting Huang, *Eyewitness Memory Contamination Through Misleading Questions by Reporters*, 42 News Rsch. J. 346 (2021), https://doi.org/10.1177/07395329211030628.

179. Morris Goldsmith, Asher Koriat & Amit Weinberg-Eliezer, *Strategic Regulation of Grain Size Memory Reporting*, 131 J. Exper. Psych.: Gen. 73 (2002), https://doi.org/10.1037//0096-3445.131.1.73; Nathan Weber & Neil Brewer, *Eyewitness Recall: Regulation of Grain Size and the Role of Confidence*, 14 J. Exper. Psych.: Applied 50 (2008), https://doi.org/10.1037/1076-898x.14.1.50.

180. Muhammad Mussaffa Butt, et. al, *Eyewitness Memory in the News Can Affect the Strategic Regulation of Memory Reporting*, 30 Memory 763 (2022), https://doi.org/10.1080/09658211.2020.1846750.

"Pristine" conditions for eyewitness interrogation are often sought in order to avoid these forms of social biasing,[181] much in the way that restricted access to task-irrelevant information is intended to avoid bias in forensic analyses.[182] Pristine is good and aspirational, of course, but in the real world it becomes very difficult to limit access to information and defuse explicit or implicit social pressures.[183]

The long reach of modern social networks

The social networks of yore, which may have included school and work colleagues, or friends from the neighborhood, have become overshadowed today by vast online communities. Active participation in these communities frequently involves presenting yourself for validation, but the reflection is from a "trick mirror,"[184] which distorts who you are—often without your awareness—and makes powerful social demands on who you should be, what you should believe, and how you should behave.[185] The rewards for conformity are addictive.[186] Naturally, this form of social influence has profound implications for the validity of eyewitness identification, such that in the extreme, "some eyewitness identifications may not be governed by memory at all."[187] Social pressures from online communities may cause a witness to question what they actually saw or, in the worst case, modify their memories without awareness. The "Rashomon effect" is the archetype of this phenomenon,[188] but we now see this happening on a massive scale in which scores of people collectively succumb to social distortions of reality and modify their criteria for judgment.[189]

181. Wixted & Wells, *supra* note 25.

182. Nat'l Comm. Forensic Sci., Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information (2015), https://perma.cc/DT64-JLGM.

183. *A National Survey of Eyewitness Identification*, *supra* note 11.

184. Jia Tolentino, Trick Mirror: Reflections on Self-Delusion (2019).

185. Jonathan Haidt, The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness (2024).

186. *Id.*

187. Margaret Bull Kovera & Andrew J. Evelo, *Eyewitness Identification in its Social Context*, 10 J. Applied Rsch. Memory & Cognition 313 (2021), https://doi.org/10.1016/j.jarmac.2021.04.003.

188. *Rashomon Effect*, Wikipedia, https://perma.cc/DYM4-8T86.

189. Dustin P. Calvillo, Justin D. Harris & Whitney C. Hawkins, *Partisan Bias in False Memories for Misinformation About the 2021 U.S. Capitol Riot*, 31 Memory 137 (2023), https://doi.org/10.1080/09658211.2022.2127771; Seth Oranburg, *Social Media and Democracy After the Capitol Riot, or, a Cautionary Tale of the Giant Goldfish*, 73 Mercer L. Rev. 591 (2021); Brenna M. Davidson & Tetsuro Kobayashi, *The Effect of Message Modality on Memory for Political Disinformation: Lessons from the 2021 U.S. Capitol Riots*, 132 Comput. Hum. Behav. 107241 (2022), https://doi.org/10.1016/j.chb.2022.107241.

Large-scale social influences on perception and judgment are not new; the problem is simply exacerbated by online social media. One of the most extraordinary non–social media examples in recent history is the eyewitness report by thousands of devout Portuguese Catholics that on

In the end, there may be little that can be practically done beyond basic education on critical thinking to prevent large-scale social influences on memory and judgment. But law enforcement, the judiciary, juries, and witnesses themselves must be made aware of these growing risks in order to evaluate the probability that eyewitness testimony is accurate.

Evidence-Based Suspicion

The concern here is that a person may be placed as a suspect in a police lineup in the absence of any plausible evidence that the person had anything to do with the crime in question. There are no legal constraints that would prevent this practice. While there are few data that bear on the frequency with which it occurs, evidence indicates that it is not uncommon.[190] In fact, a recent survey of U.S. law enforcement agencies revealed that a significant fraction operate on the belief that they need no evidence at all, or perhaps little more than a hunch, before placing a person in a lineup.[191] In 2020, the American Psychology-Law Society (AP-LS) charged six experts with developing modern policy and procedure recommendations for eyewitness identifications. Prominent among the recommendations in the AP-LS Report is the law enforcement use of "evidence-based suspicion" as a minimum criterion for lineup construction:

> There should be evidence-based grounds to suspect that an individual is guilty of the specific crime being investigated before including that individual in an identification procedure and that evidence should be documented in writing prior to the lineup.[192]

October 13, 1917, the midday "sun's disc did not remain immobile. This was not the sparkling of a heavenly body, for it spun round on itself in a mad whirl when suddenly a clamor was heard from all the people. The sun, whirling, seemed to loosen itself from the firmament and advance threateningly upon the earth as if to crush us with its huge fiery weight." John De Marchi, The Immaculate Heart: The True Story of Our Lady of Fátima (1952) (quote attributed to eyewitness Almeida Garrett). Astronomical records indicate that this "Miracle of the Sun" did not happen, but that is not what the witnesses remembered, reported, and incorporated into life decisions. *See also* Jeffrey S. Bennett, When the Sun Danced: Myth, Miracles, and Modernity in Early Twentieth-Century Portugal (2012).

190. John T. Wixted et al., *Estimating the Reliability of Eyewitness Identifications from Police Lineups*, 113 PNAS 304 (2016), https://doi.org/10.1073/pnas.1516814112; Bruce W. Behrman & Regina E. Richards, *Suspect/Foil Identification in Actual Crimes and in the Laboratory: A Reality Monitoring Analysis*, 29 Law & Hum. Behav. 279 (2005), https://doi.org/10.1007/s10979-005-3617-y.

191. Richard A. Wise, Martin A. Safer & Christina M. Maro, *What U.S. Law Enforcement Officers Know and Believe About Eyewitness Factors, Eyewitness Interviews and Identification Procedures*, 25 Applied Cognitive Psych. 488 (2011), https://doi.org/10.1002/acp.1717

192. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

This recommendation,[193] focused on law enforcement, is founded on the obvious but often overlooked fact that the simple act of placing someone in a lineup markedly increases the likelihood that they will be accused and prosecuted for a crime. Moreover, if no preexisting evidence ties that person to the crime in question, then the practice will naturally increase the probability that an innocent person is charged. There is also the worrisome question of what underlies a decision by law enforcement to include someone in a lineup in the absence of evidence for suspicion. Hunches of this sort may stem, in part, from the same prejudices that underlie biases in police traffic stops.[194] They may also arise from demographic information, including where a person lives relative to the crime scene, prior criminal history, or baseless anonymous tips, none of which rise to a level that could implicate that person, to the exclusion of others, as a legitimate suspect.

The resulting AP-LS recommendation that placement of a person in a lineup must be made only with "evidence-based suspicion" is thus a powerful strategy to avoid wrongful prosecution and conviction.[195] While this is a preemptive strategy that is necessarily under the control of the police, not the courts, knowledge of the practice used in any particular case can assist the court in evaluating the probability that the eyewitness testimony is correct.

Prior Exposure and Transference Effects

One feature of memory retrieval highlighted above under basic research (see section titled "Memory Storage" above) is the phenomenon of source memory failure and interference, in which prior knowledge from an "incorrect" source has been incorporated into a retrieved memory, which then biases memory-based decisions and impairs object recognition. This can happen in the context of eyewitness identification procedures employed by police if a witness has a prior opportunity to see a face that is part of a subsequent lineup. In the simplest case, the witness may be shown a mug shot of a potential suspect prior to the actual conduct of a lineup. The mug-shot face is then familiar to the witness at the time of a lineup that includes the face. But the witness may have lost all knowledge of why the face is familiar and naturally attributes it to their memory of the witnessed crime and the perpetrator. In simple terms, the critical memory

193. The recommendation was also made earlier by Gary Wells as the "reasonable suspicion criterion." Gary L. Wells, *Eyewitness Identification: Systemic Reforms*, 2006 Wis. L. Rev. 615 (2006).

194. Emma Pierson et al., *A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States*, 4 Nature Hum. Behav. 736 (2020), https://doi.org/10.1038/s41562-020-0858-1.

195. Failure to follow this recommendation would also seem to violate Fourth Amendment protection against unreasonable searches and seizures.

has been "contaminated," without awareness, or it has been flat-out replaced by the memory of the mug-shot photo (a phenomenon termed transference).[196] The expected result, which has been well documented in applied eyewitness studies, is that identification decisions made under conditions of unconscious memory contamination or transference are frequently incorrect, and often made with inflated confidence.[197] Not surprisingly, the same effects can be elicited by showing a witness multiple lineups that contain the same suspect.[198]

Of critical importance here is the need for the court to be aware of all prior identification procedures—a recommendation made by the NRC committee on eyewitness evidence,[199] and by authors of the recent AP-LS document on eyewitness policy and procedures[200]—as this information bears on the probability that the eyewitness testimony is correct. We note here that the 1968 Supreme Court, *Simmons v. United States*, fully appreciated the concern about memory contamination from photos and the implications for accuracy of testimony.[201]

196. Jennifer E. Dysart et al., *Mug Shot Exposure Prior to Lineup Identification: Interference, Transference, and Commitment Effects*, 86 J. Applied Psych. 1280 (2001), https://doi.org/10.1037//0021-9010.86.6.1280.

197. Committee Report, Report to the Secretary of State for the Home Department of the Departmental Committee on Evidence of Identification in Criminal Cases (1976) [hereinafter Devlin Report]; Evan Brown, Kenneth A. Deffenbacher & William Sturgill, *Memory for Faces and the Circumstances of Encounter*, 62 J. Applied Psych. 311 (1977), https://doi.org/10.1037/e666602011-038; Kenneth A. Deffenbacher, Brian H. Bornstein & Steven D. Penrod, *Mugshot Exposure Effects: Retroactive Interference, Mugshot Commitment, Source Confusion, and Unconscious Transference*, 30 Law & Hum. Behav. 287 (2006), https://doi.org/10.1007/s10979-006-9008-1; Dysart et al., *supra* note 196; Graham Davies, John Shepherd & Hadyn Ellis, *Effects of Interpolated Mugshot Exposure on Accuracy of Eyewitness Identification*, 64 J. Applied Psych. 232 (1979), https://doi.org/10.1037//0021-9010.64.2.232.

198. Steblay et al., *supra* note 157.

199. Identifying the Culprit, *supra* note 12.

200. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

201. Simmons v. United States, 390 U.S. 377, 383–84 (1968) ("It must be recognized that improper employment of photographs by police may sometimes cause witnesses to err in identifying criminals. A witness may have obtained only a brief glimpse of a criminal, or may have seen him under poor conditions. Even if the police subsequently follow the most correct photographic identification procedures and show him the pictures of a number of individuals without indicating whom they suspect, there is some danger that the witness may make an incorrect identification. This danger will be increased if the police display to the witness only the picture of a single individual who generally resembles the person he saw, or if they show him the pictures of several persons among which the photograph of a single such individual recurs or is in some way emphasized. The chance of misidentification is also heightened if the police indicate to the witness that they have other evidence that one of the persons pictured committed the crime. Regardless of how the initial misidentification comes about, the witness thereafter is apt to retain in his memory the image of the photograph, rather than of the person actually seen, reducing the trustworthiness of subsequent lineup or courtroom identification.").

Identification Procedures

Law enforcement can readily adopt procedures for how they present faces to an eyewitness. Few topics in applied eyewitness research have spawned as much scientific experimentation, debate, and impact on law enforcement practices, as how to best design these identification procedures. While the goal of these efforts has been to prospectively improve eyewitness performance to better serve the needs of criminal investigation and prosecution, knowledge of the identification procedures that were used in a given case can assist courts in retrospective evaluation of the accuracy of testimony.

Showups

In the simplest identification task, known as a "showup," the suspect alone is presented to the witness, who responds yes or no. One common reason for conducting a showup is convenience and exigence: Shortly after arriving at the crime scene, the police apprehend a suspect fitting the witness's description, who is then shown to the witness for confirmation. Despite the seeming benefits of this procedure, it is flawed, owing to suggestive qualities that may lower a witness's criterion for an identification: (1) the witness's knowledge that the suspect was apprehended nearby sets up an expectation, (2) the presentation of a sole suspect leads to an inference that the police must be quite certain, and (3) the impromptu context encourages the witness to be forthcoming with assistance. Under the circumstances, we should naturally expect that showups will increase the probability of correct identification (it's much easier to identify your suitcase if it is the only one on the carousel), but they should also increase the probability of making an incorrect identification. Not surprisingly, applied eyewitness studies demonstrate that mistaken identifications are more common using the showup procedure than with "standard" six-person lineups,[202] particularly when innocent suspects resemble the witness's description of the perpetrator.[203]

Confirmatory photo identification

Police will, on occasion, display a single photograph to a witness in an effort to confirm the identity of a person of interest. Police typically limit this method

202. A. Daniel Yarmey, Meagan J. Yarmey & Linda A. Yarmey, *Accuracy of Eyewitness Identifications in Showups and Lineups*, 20 Law & Hum. Behav. 459 (1996), https://doi.org/10.1007/bf01498981.

203. Nancy Steblay et al., *Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta–Analytic Comparison*, 27 Law & Hum. Behav. 523, 523–40 (2003), https://doi.org/10.1023/a:1025438223608.

to situations in which the person is previously known to or acquainted with the witness. The goal is to clarify the legal identity (birth name/government name) of an individual who is well known to a witness, but only by a street name. In such examples, a witness may know (and may have known) the perpetrator for years but may be able to identify him only by a street name, such as "Diamond." The police will typically use this procedure before making an arrest. As for mug books and showups, the procedure is highly suggestive, as it implicates the person in the photo, which may, in turn, unconsciously contaminate or replace the witness's memory of the face of the perpetrator.[204] The practical consequence of this procedure is spelled out above in the section titled "Prior Exposure and Transference Effects," namely reduced identification accuracy in a subsequent lineup, and artifactually elevated confidence on the part of the witness.[205] The existence of this practice highlights, once again, the need for courts to have knowledge of all prior identification procedures in a given case in order to evaluate their implications for eyewitness accuracy.[206]

Lineups: Simultaneous versus sequential procedures

To engage discriminative abilities and limit selection bias, multiperson lineups have been the standard since at least the nineteenth century. The people in the lineup commonly include the suspect and a set of people known as "fillers," who are known to be innocent and (ideally) appear similar to the witness's description of the perpetrator. In laboratory studies, "target absent" (innocent suspect) lineups are also used to enable a measure of the probability of identifying the known culprit relative to the probability of identifying an innocent person. In many studies conducted since the 1970s, this target present versus target absent comparison has been the standard approach to empirically evaluate the effects of estimator and system variables on eyewitness performance (many of which are cited in the foregoing sections).

In the 1980s, the discussion of factors that affect eyewitness performance broadened to include the nature of the multiperson lineup itself. Lindsay and Wells hypothesized that simultaneous viewing of lineup participants—the standard practice of the time—was a factor that prejudiced identifications.[207] In particular, they posited that visual comparisons between faces (relative comparisons) could take precedence over the comparisons that really mattered—that is,

204. Wells et al., *Recommendations for Lineups and Photospreads, supra* note 27, at 457–70.

205. Deffenbacher et al., *Mugshot Exposure Effects, supra* note 197, at 287–307.

206. Identifying the Culprit, *supra* note 12; Devlin Report, *supra* note 197, at 3.

207. R.C. Lindsay & Gary L. Wells, *Improving Eyewitness Identifications from Lineups: Simultaneous Versus Sequential Lineup Presentation*, 70 J. Applied Psych. 556 (1985), https://doi.org/10.1037//0021-9010.70.3.556.

those between each lineup face and the visual memory of the perpetrator (absolute comparisons), leading to a selective increase in identifications of innocent suspects. To overcome this assumed problem, Lindsay and Wells proposed a novel lineup in which participants are viewed sequentially.

Early laboratory comparisons of performance on simultaneous versus sequential lineups appeared to demonstrate a sequential advantage, based on a higher ratio of correct to incorrect identifications.[208] In lockstep with this science and mindful of wrongful convictions, many jurisdictions adopted the new sequential procedure.[209] The problem with this outcome originates with the measure of performance—the diagnosticity ratio (DR)—which is affected both by the eyewitness's ability to discriminate the perpetrator from an innocent suspect and by the criterion used to make an identification decision. It could have been true that sequential lineups also improve discriminability, which would be of real value, but it is impossible to know that using a single DR measure.

There is an established procedure for evaluating discriminability independently of the decision criterion. This procedure, known as analysis of receiver operating characteristics (ROC), is drawn from signal detection theory.[210] Figure 13 summarizes results from over 20 years of studies that have employed the ROC approach. Each dot represents data from a published study, identified at right. The x-axis plots a measure of the difference between discriminability measures obtained for the two lineup types, with positive values (SIM > SEQ) indicating a simultaneous advantage, and negative values (SEQ > SIM) indicating a sequential advantage. (Cohen's *d* is a standard index of effect size.) Dot size reflects sample size (the number of people tested in the study); the y-axis orders studies as a function of sample size. Green dots indicate studies for which the discriminability difference for lineup types is statistically significant. Thus, contrary to initial claims for the superiority of the sequential lineup based on the DR measure, collective evidence supports a small but consistent improvement in discriminability for the simultaneous lineup relative to sequential.[211] Extrapolation from the laboratory suggests that identifications made in jurisdictions that employ the sequential procedure may not be based on the best human discriminative abilities, which is a factor for consideration when assessing the probability that eyewitness testimony is accurate.

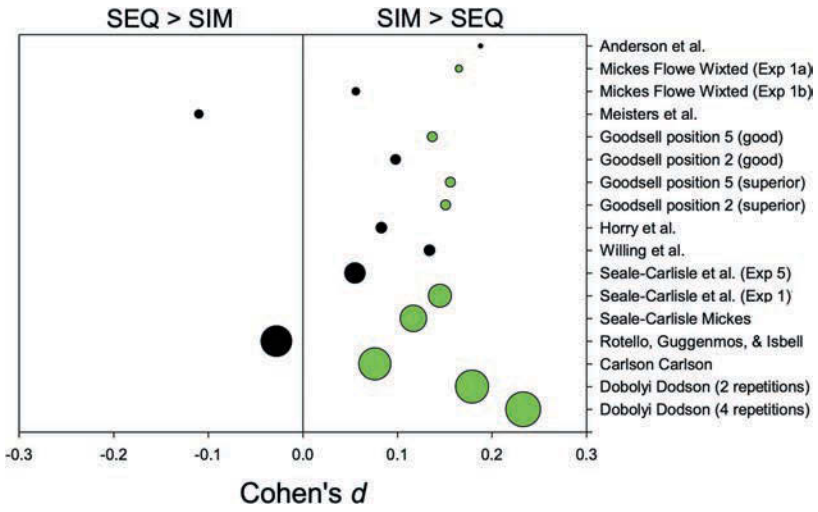
208. *Id.*

209. A National Survey of Eyewitness Identification Procedures, *supra* note 11.

210. Signal detection theory includes a framework for differentiating a person's ability to discriminate the presence and absence of a stimulus from the criterion the person uses to make that decision. David M. Green & John A. Swets, Signal Detection Theory and Psychophysics (1966).

211. Travis M. Seale-Carlisle et al., *Designing Police Lineups to Maximize Memory Performance*, 25 J. Exper. Psych.: Applied 410 (2019), https://doi.org/10.1037/xap0000222; Laura Mickes & John T. Wixted, *Eyewitness Memory*, *in* 2294 Oxford Handbook of Human Memory (2022).

Figure 13. Summary of empirical support for simultaneous lineup advantage.



Source: Adapted from Laura Mickes & John T. Wixted, *Eyewitness Memory, in* Oxford Handbook of Human Memory (2022) and used with permission of Oxford Publishing Limited through PLSclear. Data sources: Shannon M. Andersen et al., *Individual Differences Predict Eyewitness Identification Performance*, 60 Personality & Individual Differences, 36 (2014), https://doi.org/10.1016/j.paid.2013.12.011; Laura Mickes et al., *Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous Versus Sequential Lineups*, 18 J. Experimental Psych.: Applied, 361 (2012), https://doi.org/10.1037/a0030609; Julia Meisters et al., *Eyewitness Identification in Simultaneous and Sequential Lineups: An Investigation of Position Effects Using Receiver Operating Characteristics*, 26 Memory, 1297 (2018), https://doi.org/10.1080/09658211.2018.1464581; Charles A. Goodsell, *Effects of Eyewitness Memory Encoding Strength on Sequential and Simultaneous Lineup Identifications*, (unpublished manuscript) (on file with Department of Psychology, Canisius College, Buffalo, New York) (2019); Ruth Horry et al., *The Effects of Allowing a Second Sequential Lineup Lap on Choosing and Probative Value*, 21 Psych., Pub. Pol'y, & L., 121 (2015), https://doi.org/10.1037/law0000041; Sonja Willing et al., *Presenting a Similar Foil Prior to the Suspect Reduces the Identifiability of the Perpetrator in Sequential Lineups: A ROC-Based Analysis*, (manuscript under review) (2019); Travis M. Seale-Carlisle et al., *Designing Police Lineups to Maximize Memory Performance*, 25 J. Experimental Psych.: Applied, 410 (2019), https://doi.org/10.1037/xap0000222; Travis M. Seale-Carlisle & Laura Mickes, *US Line-ups Outperform UK Line-ups*, 3 Royal Soc'y Open Sci., 160300 (2016), https://doi.org/10.1098/rsos.160300; Caren Rotello et al. (unpublished data); Curt A. Carlson & Maria A. Carlson, *An Evaluation of Perpetrator Distinctiveness, Weapon Presence, and Lineup Presentation Using ROC Analysis*, 3 J. Applied Rsch. Memory & Cognition, 45 (2014), https://doi.org/10.1016/j.jarmac.2014.03.004; David G. Dobolyi & Chad S. Dodson, *Eyewitness Confidence in Simultaneous and Sequential Lineups: A Criterion Shift Account for Sequential Mistaken Identification Overconfidence*, 19 J. Experimental Psych.: Applied, 345 (2013), https://doi.org/10.1037/a0034596.

Other lineup procedures

One of the limitations of traditional lineup procedures is that eyewitnesses are asked to make two distinct contributions to the decision: (1) measurement of the similarity between memory and the face of a lineup participant, and (2) classification of the face as a match or nonmatch. An eyewitness identification manifests the classification decision, but the recipient of that decision has no direct access to the measurement upon which it was based, or to the criterion that the eyewitness used to evaluate the measurement, leaving the decision susceptible to unrecognized bias. This is notably distinct from man-made information processing devices (e.g., smartphone fingerprint and face detectors), which yield similarity measurements that are accessible and classifiable by an objective statistical procedure. A lineup procedure could, in principle, render a similarity measure but exclude the eyewitness from the classification—the latter performed instead by the lineup administrator using an objective statistical analysis. Recent laboratory studies demonstrate that this is a feasible, high-accuracy approach that could transform the utility of eyewitness lineups,[212] but adoption by the criminal justice system must wait for evaluation in field studies with real casework, as we consider below.

In-court identifications

In criminal trials that are built on eyewitness testimony, a nearly ubiquitous feature is the theatrical event in which a witness is asked to identify the perpetrator in front of the jury. This practice has two consequences that adversely impact efforts by jurors to assess the probability that the witness testimony is correct.

First, the witness's level of confidence in the identification made at trial is generally highly inflated. In the Courtney case summarized above (see section titled "An Illustrative Case"), the victim-witness reported a confidence level of 60% upon initial identification. After identifying Courtney in court at the request of the prosecution, the witness expressed their confidence as: "I will never forget what he looks like." The section titled "Timing of identification and conditions of confidence assessment," above, delves into the specific reasons why this happens.

Second, the inflated confidence expressed during an in-court identification is not commonly recognized as such by a lay jury. It thus becomes a very powerful form of rhetoric that conveys authority on the part of the witness and stirs the

212. Sergei Gepshtein et al., *A Perceptual Scaling Approach to Eyewitness Identification*, 11 Nature Commc'ns 3380 (2020), https://doi.org/10.1038/s41467-020-17194-5; Sergei Gepshtein & Thomas D. Albright, *Thinking Outside the Lineup Box: Eyewitness Identification by Perceptual Scaling*, 10 J. Applied Rsch. Memory & Cognition 221 (2021), https://doi.org/10.1016/j.jarmac.2021.05.001.

emotions of the jurors.[213] All of which moves them toward a decision to convict.[214] As Supreme Court Justice William Brennan wrote, "[T]here is almost nothing more convincing than a live human being who takes the stand, points a finger at the defendant, and says 'That's the one!'"[215] This "convincing," of course, is the primary reason why expert testimony or carefully worded jury instructions can be valuable, particularly when they explain the reasons *why* confidence is inflated.

In view of the potential for biasing the outcome, one might argue that in-court identifications should be eliminated altogether in criminal trials.[216] This proposal has gained some traction. The Massachusetts Supreme Judicial Court has ruled, for example, that in-court identifications should not normally be permitted unless they are the first identification, or if the out-of-court identification was suppressed.[217] The Connecticut Supreme Court has ruled to allow an in-court identification only if the witness knew the defendant before the crime, or if a prior lineup was proved to be nonsuggestive.[218] These rules have exceptions, as noted, and other courts conduct a burden-shifting approach toward in-court identifications; an in-court identification could also potentially be so unduly suggestive as to amount to a violation of *Manson v. Brathwaite*.[219] Research supports the view that in-court identifications should be handled with great caution, and that jurors must be made aware that they should not place such great weight on the in-court confidence of an eyewitness.

213. Maksymilian Del Mar, *The Confluence of Rhetoric and Emotion: How the History of Rhetoric Illuminates the Theoretical Importance of Emotion*, 36 Law & Literature 1 (2024), https://doi.org/10.1080/1535685x.2022.2115651; Aristotle, The Rhetoric of Aristotle: A Translation (Richard Claverhouse Jebb trans., 1909).

214. Brewer & Burke, *supra* note 173; Slane & Dodson, *supra* note 173.

215. Watkins v. Sowders, 449 U.S. 341, 353 (1981) (Brennan, J., dissenting) (citing Elizabeth F. Loftus, Eyewitness Testimony (1979)).

216. Brandon L. Garrett, *Eyewitnesses and Exclusion*, 65 Vand. L. Rev. 451 (2012).

217. *See* Commonwealth v. Johnson, 45 N.E.3d 83, 92–94 (Mass. 2016) (reasoning that "a subsequent in-court identification cannot be more reliable than the earlier out-of-court identification, given the inherent suggestiveness of in-court identifications and the passage of time"); Commonwealth v. Crayton, 21 N.E.3d 157, 169 (Mass. 2014) ("Where an eyewitness has not participated before trial in an identification procedure, we shall treat the in-court identification as an in-court showup, and shall admit it in evidence only where there is 'good reason' for its admission.").

218. State v. Dickson, 141 A.3d 810, 817 (Conn. 2016) ("[I]n cases in which identity is an issue, in-court identifications that are not preceded by a successful identification in a nonsuggestive identification procedure implicate due process principles and, therefore, must be prescreened by the trial court."). Note that this ruling has been challenged. *See* Tatum v. Comm'r of Correction, No. 20727, 2024 WL 3433265, at *3 (Conn. July 16, 2024).

219. *See* State v. Hickman, 330 P.3d 551, 568 (Or. 2014) (en banc) ("Courts considering the admissibility of first-time in-court identifications generally have placed the burden of seeking a prophylactic remedy on the defendant."); United States v. Domina, 784 F.2d 1361, 1369 (9th Cir. 1986) (noting that district court's denial of request for in-court lineup will only be overturned if in-court identification procedures were so suggestive and conducive to irreparable misidentification as to amount to denial of due process).

Filler Selection

Fillers are lineup participants known to be innocent, who serve as lures to challenge recognition memory. The choice of fillers has long been known to markedly influence eyewitness performance in laboratory studies.[220] People recognize objects probabilistically based on the degree to which they elicit a memory signal corresponding to a particular target previously seen. It naturally follows that similar objects are more likely to elicit the same memory signal and thus have similar likelihoods of being recognized as the target. The similarity of fillers to the witness's description of the perpetrator is thus an important variable that affects both eyewitness discriminability and accuracy.

Lineups composed of fillers who are all of roughly the same degree of similarity to the witness's description are termed "fair." Conversely, lineups composed of fillers that possess differing degrees of similarity to the description are termed "unfair" or "biased." An unfair lineup, in which one filler is closer in similarity to the description of the perpetrator, reduces uncertainty by reducing the number of sensible choices, thus increasing the likelihood of correctly identifying the perpetrator, but also increasing the likelihood of misidentifying someone who looks like the perpetrator.[221]

Despite the well-understood and potentially disastrous consequences of unfair lineups, there have been few attempts to systematize the law enforcement process of filler selection. Published guidelines for filler selection state that lineups should be constructed to ensure that "the suspect does not unduly stand out" and lineups should "avoid using fillers that so closely resemble the suspect that a person familiar with the suspect might find it difficult to distinguish the suspect from the fillers."[222] This guidance—fillers should be similar to the suspect but not too much so—is open to interpretation and is applied by different agents in different ways. It has not been effective.[223]

One emerging approach would use face similarity metrics to create lineups in which fillers are of known similarity to each other and to the witness's description of the perpetrator. A recent meta-analysis of laboratory studies adopting this approach revealed that decreasing the similarity between the suspect and fillers has the effect of increasing the probability that the suspect is identified,

220. Roy S. Malpass & Patricia G. Devine, *Measuring the Fairness of Eyewitness Identification Lineups*, *in* 81 Evaluating Witness Evidence (1983); Gary L. Wells, Michael R. Leippe & Thomas M. Ostrom, *Guidelines for Empirically Assessing the Fairness of a Lineup*, 3 Law & Hum. Behav. 285 (1979), https://doi.org/10.1007/bf01039807.

221. This effect is much like the consequence of a showup identification.

222. U.S. Dep't of Just., Nat'l Inst. of Just., *supra* note 27.

223. Carmen A. Lucas, Neil Brewer & Matthew A. Palmer, *Eyewitness Identification: The Complex Issue of Suspect-Filler Similarity*, 27(2) Psych., Pub. Pol'y, & L. 151 (2021).

regardless of whether that identification is correct.[224] The case of Uriah Courtney, highlighted above, illustrates the tragic consequences of this practice in the real world: Courtney's conviction was based on two witness identifications using a lineup in which Courtney stood out from many of the fillers as an exemplar of the witness descriptions.[225] By contrast, increasing the similarity between suspect and fillers increases the probability of filler identification or a lineup rejection, meaning that fewer suspects are identified overall. A more recent analysis suggests that this is a consequence of reduced discriminability between perpetrator and innocent suspect.[226]

A recent laboratory study adds a new wrinkle to this picture and a promising target for future investigation: Testing the counterintuitive predictions of a signal detection model, the investigators found that "choosing fillers who match the description of the perpetrator but who are otherwise dissimilar to the suspect's appearance yielded fair lineups and enhanced eyewitness identification performance."[227] To appreciate this conclusion, suppose that the witness's verbal description of the perpetrator included red hair and a goatee. A suspect with those features is apprehended, and all lineup fillers are selected to have the same diagnostic features (i.e., the lineup is fair). The perpetrator may have other features not mentioned by the witness, such as blue eyes. Fillers may share the blue eyes feature with the perpetrator by chance, increasing the probability that the witness will misidentify them. Thus, the lineup is more likely to yield a correct outcome if fillers are chosen without regard to features that were not part of the witness's description of the perpetrator.

Recent evidence points to another variable associated with filler–suspect similarity that has a somewhat counterintuitive effect on eyewitness performance: the size of the face database from which fillers are selected. One might expect that having a large library of faces to draw from, such as computer databases of driver's license photos, would make it possible to find filler faces that are more suitable for a lineup and may facilitate eyewitness performance.[228] Contrary to this intuition, selection of fillers from very large databases is commonly associated with greater similarity between the suspect and lineup fillers, which has the consequence of increasing the frequency of filler picks by eyewitnesses, thus reducing accuracy.[229]

224. Ryan J. Fitzgerald et al., *The Effect of Suspect-Filler Similarity on Eyewitness Identification Decisions: A Meta-Analysis*, 19 Psych., Pub. Pol'y, & L. 151 (2013), https://doi.org/10.1037/a0030618.

225. Albright, *Why Eyewitnesses Fail*, *supra* note 14, at 7758–64.

226. Fitzgerald et al., *supra* note 224, at 151.

227. Melissa F. Colloff et al., *Optimizing the Selection of Fillers in Police Lineups*, 118 PNAS e2017292118 (2021), https://doi.org/10.1073/pnas.2017292118.

228. Fitzgerald et al., *supra* note 224, at 151.

229. Amanda N. Bergold & Paul Heaton, *Does Filler Database Size Influence Identification Accuracy?*, 42 Law & Hum. Behav. 227 (2018), https://doi.org/10.1037/lhb0000289.

With a movement toward automated systems for filler selection, their ease of use, and their adoption by the police,[230] database size is a critical variable.

There are many more details to be worked out in this space of filler–suspect similarity, but on the whole, it is clear that filler selection matters greatly to eyewitness performance, and thus the approach that was taken in a given case may offer insight into the probability that the eyewitness testimony is correct.

Corroborating Variables: Machine-Based Facial Recognition

There are, of course, many other variables associated with eyewitness events that fall in neither the estimator nor system category. For police investigations and criminal trials involving an eyewitness, the most useful of these variables are sources of forensic evidence that bear on, and potentially corroborate, eyewitness identification. The forms of evidence in this category are numerous, including, but not limited to, fingerprints, tool marks, chemical/biological residues, and DNA. The utility of these forms of evidence for the courts is treated elsewhere.[231]

One specific and very recent form of corroborating evidence deserves mention here: automated facial recognition. New artificial intelligence (AI) technology for machine-based person recognition derived from photographs or video collected in the real world, combined with the dramatic proliferation of live cameras in public spaces, has enabled law enforcement to harvest identification evidence that may complement traditional eyewitness testimony. This is a rapidly evolving and largely unregulated field, which has considerable relevance to judicial decisions about the accuracy of eyewitness testimony from real people. We refer readers to a report from the National Academies of Sciences, Engineering, and Medicine, which details the technology and its accuracy, utility, potential applications, and privacy concerns.[232]

Applied Eyewitness Studies in the Field

The results of applied studies conducted in the laboratory have revealed a large number of variables that affect the ability of witnesses to discriminate between the perpetrator and innocent suspects, as well as the overall accuracy of

230. A National Survey of Eyewitness Identification Procedures, *supra* note 11.

231. *See, e.g.*, Valena E. Beety, Jane Campbell Moriarty & Andrea L. Roth, *Reference Guide on Forensic Feature Comparison Evidence*, in this manual.

232. Nat'l Acads. Scis., Eng'g & Med., Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance (2024), https://doi.org/10.17226/27397.

identifications made. These findings both encourage this empirical strategy and hold promise for reform, but they have often been criticized for lack of ecological validity.[233] Specifically, the laboratory crime is a recognized pretense that cannot be expected to elicit critical states of the observer, such as high values of stress, fear, pain, and loss of attentional focus. The ultimate test of the utility of laboratory discoveries is thus successful translation to real criminal casework as evaluated by field studies.[234]

There are three constraints associated with achieving this translational goal. First, it requires access to a high-functioning police department with a willingness to test scientific hypotheses. Second, ground truth is unknown in real cases, which means that experimental efforts must rely on independent corroboration of truth ("extrinsic evidence") as a means to validate the effects of the variables tested. That reliable corroboration is often difficult to come by. Third, the situational context in the field is difficult to control with the level of precision that is common in a laboratory. Real-world things sometimes happen unexpectedly, or there may simply be no record of what did happen.

Operating under these constraints, a small number of studies have tested the effects of estimator and system variables in the field. Results generally support findings from laboratory studies of effects of retention interval, cross-race effect, weapon focus, and lineup type,[235] but many factors have been poorly

233. Yoojin Chae, *Application of Laboratory Research on Eyewitness Testimony*, 10 J. Forensic Psych. Prac. 252 (2010), https://doi.org/10.1080/15228930903550608; Graham F. Wagstaff et al., *Can Laboratory Findings on Eyewitness Testimony Be Generalized to the Real World? An Archival Analysis of the Influence of Violence, Weapon Presence, and Age on Eyewitness Accuracy*, 137 J. Psych. 17 (2003), https://doi.org/10.1080/00223980309600596.

234. Daniel L. Schacter et al., *Policy Forum: Studying Eyewitness Investigations in the Field*, 32 Law & Hum. Behav. 3 (2008), https://doi.org/10.1007/s10979-007-9093-9.

235. John C. Yuille & Judith L. Cutshall, *A Case Study of Eyewitness Memory of a Crime*, 71 J. Applied Psych. 291 (1986), https://doi.org/10.1037//0021-9010.71.2.291; Yarmey et al., *supra* note 202, at 459–77; Bruce W. Behrman & Sherrie L. Davey, *Eyewitness Identification in Actual Criminal Cases: An Archival Analysis*, 25 Law & Hum. Behav. 475 (2001), https://doi.org/10.1023/a:1012840831846; Amy Klobuchar, Nancy K. Mehrkens Steblay & Hilary Lindell Caligiuri, *Improving Eyewitness Identifications: Hennepin County's Blind Sequential Lineup Pilot Project*, 4 Cardozo Pub. L., Pol'y & Ethics J. 381 (2006); Daniel L. Schacter et al., *Policy Forum: Studying Eyewitness Investigations in the Field*, 32 L. & Hum. Behav. 3 (2008), https://doi.org/10.1007/s10979-007-9093-9; Nancy K. Steblay, *What We Know Now: The Evanston Illinois Field Lineups*, 35 Law & Hum. Behav. 1 (2011), https://doi.org/10.1007/s10979-009-9207-7; Gary L. Wells, Nancy K. Steblay & Jennifer E. Dysart, *A Test of the Simultaneous vs. Sequential Lineup Methods: An Initial Report of the AJS National Eyewitness Identification Field Studies*, Am. Judicature Soc. (2011), https://perma.cc/42F5-LVL5; Gary L. Wells, Nancy K. Steblay & Jennifer E. Dysart, *Double-Blind Photo Lineups Using Actual Eyewitnesses: An Experimental Test of a Sequential Versus Simultaneous Lineup Procedure*, 39 Law & Hum. Behav. 1 (2015), https://doi.org/10.1037/lhb0000096; Karen L. Amendola & John T. Wixted, *Comparing the Diagnostic Accuracy of Suspect Identifications Made by Actual Eyewitnesses from Simultaneous and Sequential Lineups in a Randomized Field Trial*, 11 J. Exper. Crim. 263 (2015), https://doi.org/10.1007/s11292-014-9219-2; Wixted et al., *supra* note 190.

controlled. Moreover, in real police investigations it can be hard to assess ground truth, and cognitive biases can play a complicating role in perceptions of evidence strength. Much more needs to be done in this difficult arena before changes are made to policy and practice. It seems possible that the laboratory versus field relationship will break down where the real emotion of criminal witnessing comes into play.

Effects of Eyewitness Testimony on Juries

The identification decision made by an eyewitness is only the first stage in which human factors may drive a criminal prosecution off track. The second stage is in the communication of eyewitness testimony to the trier of fact. Recent scientific studies have explored the role that communication plays in inferring the accuracy of eyewitness testimony. Specifically, the witness is viewed as the transmitter of information, and the jury is the intended receiver of that information. While much science focuses on improving the accuracy of testimony and the ability to predict that accuracy, these efforts will be in vain if the information is not correctly received by the trier of fact.

Laypeople may have a range of misconceptions regarding how eyewitness memory functions. Further, there is a traditional practice of allowing eyewitnesses to identify a person in the courtroom (see section titled "In-court identifications" above). As Justice Sonia Sotomayor explained in dissent in *Perry v. New Hampshire*: "At trial, an eyewitness' artificially inflated confidence in an identification's accuracy complicates the jury's task of assessing witness credibility and reliability."[236] Research has confirmed that these in-court identifications powerfully influence jurors, and in addition, that the confidence of an eyewitness in court has a particularly outsized influence on lay jurors.[237] An eyewitness may appear to be highly confident in court, even if the witness was highly uncertain during the earlier eyewitness identification procedure conducted by police out of court—the phenomenon known as confidence inflation (see above section titled "Timing of identification and conditions of confidence assessment").[238] Indeed, confidence can be inflated by biasing information, including feedback that eyewitnesses receive after an identification procedure and in preparation for testimony at trial.[239]

236. 565 U.S. 228, 252 (2012), (Sotomayor, J., dissenting).

237. *See generally,* Garrett et al., *supra* note 33 (detailing findings of mock juror survey regarding weight given to confidence of eyewitnesses).

238. *See* Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 21.

239. Carl Martin Allwood, Jens Knutsonn & Pär Anders Granhag, *Eyewitnesses Under Influence: How Feedback Affects the Realism in Confidence Judgements*, 12 Psych. Crime & L. 25, 36 (2006), https://doi.org/10.1080/10683160512331316316.

As noted below, some jurisdictions and courts have adopted jury instructions and other mechanisms to improve communication between an eyewitness or their proxy (transmitter) and a jury (receiver), to better educate jurors regarding the limitations of eyewitness testimony, including by limiting the use of in-court identifications. There is, however, evidence that several of these more recent sets of jury instructions are not particularly effective at improving jurors' ability to discriminate between more or less reliable eyewitness identifications,[240] perhaps because they do not sufficiently address the weight that jurors place on the courtroom confidence of the eyewitness, and—as highlighted above—the rhetorical impact of an in-court identification. Indeed, recent evidence shows that more pointed instructions regarding in-court confidence, and the reasons why there are dangers of relying on it, may be more effective.[241] Expert witnesses, of course, may address these issues head-on, and with a richer description of the relevant scientific research.

Sources of Expertise on Matters of Eyewitness Science

The value and utility of eyewitness evidence lie in both the content of testimony (who did what to whom) and the probability that the testimony is correct. As we have emphasized throughout this text, there are both general and specific features of the process of perceiving criminal events, identifying perpetrators, and testifying to experience, that bear on the probability that testimony is accurate. Knowledge and understanding of these features and their predictive relationship to accuracy are often beyond the ken of jurors, judges, and witnesses proffering testimony. In addition to numerous court rulings on this topic (particularly *Manson v. Brathwaite*, but also more recent federal and state court decisions that digest and rely on scientific research in the area), there are many sources of clarifying information that can help courts decide about admissibility of eyewitness testimony and assess its probative value.

These include written documents, such as (but not limited to) the present reference guide, the 2014 NRC report on eyewitness identification,[242] the 2019

240. Jones & Bergold, *supra* note 33, at 433–55; Berman, *supra* note 33; Marlee Kind Dillon et al., Henderson *Instructions: Do They Enhance Evidence Evaluation,* 17 J. Forensic Psych. Rsch. & Prac. 1 (2017); Garrett et al., *supra* note 33; Jones et al., *supra* note 33; Laub et al., *supra* note 33; Papailiou et al., *supra* note 33.

241. Brandon L. Garrett et al., *Sensitizing Jurors Regarding Eyewitness Testimony*, 12 J. Applied Rsch., Memory & Cognition 1, 141–57 (2022), https://doi.org/10.1037/mac0000035.

242. Identifying the Culprit, *supra* note 12.

report of the Third Circuit task force on eyewitness identifications,[243] and the eyewitness policy and procedural recommendations of the AP-LS.[244] The underpinning of these documents is a very large peer-reviewed scientific literature from basic and applied fields of psychology and/or cognitive and neural sciences, with particular focus on vision, memory, attention, signal detection, predictive coding, and decision-making (much of which is cited herein). In addition to these written guides, which often themselves require clarification and interpretation, decisions about admissibility of eyewitness evidence and assessment of its weight (probability that the evidence is correct) can be informed by experts with demonstrated training and proficiency in the relevant fields of science. General standards for admissibility of such experts are, of course, defined by the Supreme Court's *Daubert*[245] ruling and Rule 702 of the Federal Rules of Evidence, which emphasize both qualifications of the expert and the validity and reproducibility of the underlying science reported by the expert.[246]

On the specific topic of eyewitness evidence, expert qualifications and experience should include a published record of scholarly empirical work on phenomenology, function, and underlying mechanisms of visual perception and recognition memory, applied studies of the effects of contextual variables (viewing conditions and cognitive state of the observer) on eyewitness discriminability and accuracy, and probabilistic modeling of evidence-based behavioral choice under conditions of uncertainty. Experts with these qualifications and knowledge would commonly hold tenured or tenure-track positions at accredited universities or in nonprofit research institutes, in academic disciplines of psychology and/or cognitive and neural sciences, and statistics. Such experts can draw clear and useful (and sometimes precisely quantitative) inferences about (but not limited to) the following: (1) the effects of contextual variables on initial eyewitness collection of information about criminal events, such as atmospheric conditions (e.g., rain, fog), viewing duration, viewing distance, illuminant intensity, luminance contrast, viewing angle, image "crowding," distracting features that hijack attention, expectations, stress, and emotional states; (2) the degradation and contamination of memories of witnessed events with the passage of time; (3) the influence of particular behavioral strategies (including introduction of uncertainty and bias) for eliciting eyewitness recognition memory (e.g., lineup type, filler selection); (4) the social influence of other parties (e.g., law enforcement officers, attorneys, news media, friends, family) on choice and confidence in

243. *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 29.

244. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3–36.

245. Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

246. Thomas D. Albright, *A Scientist's Take on Scientific Evidence in the Courtroom*, 120 PNAS e2301839120 (2023), https://doi.org/10.1073/pnas.2301839120.

proffered testimony; (5) the utility of self-report measures (e.g., confidence) that may be presented as indicia of accuracy; and (6) the consequences of poor communication between eyewitness (and counsel) and trier of fact.

The Legal Framework for Eyewitness Evidence

The potential inaccuracy of eyewitness evidence has long been well known to judges. As the U.S. Supreme Court put it in its 1967 ruling in *United States v. Wade*, "The vagaries of eyewitness identification are well-known; the annals of criminal law are rife with instances of mistaken identification."[247] The traditional constitutional criminal procedure approach to such evidence has involved several related protections, including the right to counsel at an in-person lineup, as well as a due process rule requiring a deferential review of the validity of an unduly suggestive lineup. In addition, courts have conducted pretrial hearings to examine the validity of eyewitness evidence, and offered fairly basic and brief jury instructions to explain considerations relevant to eyewitness evidence to jurors. Those rules were developed in the late 1960s and 1970s, and in federal courts, the basic framework has remained largely unchanged ever since.

In the nearly five decades since, a substantial body of visual perception and eyewitness memory science, as described above, has transformed our understanding of eyewitness evidence. This research has affected police identification practices, including those of federal law enforcement agencies. In federal court, these developments have informed, for example, a move toward receptivity to expert testimony regarding the science of eyewitness vision and memory, where it would assist the jury and satisfy other requirements regarding expert testimony,[248] with other courts preferring the use of jury instructions.[249] More recently, a noteworthy Third Circuit Task Force on Eyewitness Evidence examined a range of recommendations regarding how to consider eyewitness evidence.[250]

247. United States v. Wade, 388 U.S. 218, 228 (1967).

248. *See generally,* United States v. Rodriguez-Felix, 450 F.3d 1117, 1124 (10th Cir. 2006), (summarizing "there has been a trend in recent years to allow such testimony" and "subject to the trial court's careful supervision, properly conceived expert testimony may be admissible to challenge or support eyewitness evidence" when it would assist the jury); United States v. Harris, 995 F.2d 532, 534 (4th Cir. 1993) (summarizing "[u]ntil fairly recently, most, if not all, courts excluded expert psychological testimony on the validity of eyewitness identification" but describing how courts came to recognize trial court discretion to allow such testimony).

249. *See, e.g.*, United States v. Jones, 689 F.3d 12, 20 (1st Cir. 2012) (finding it "within the district court's province to provide this information through instructions rather than through dueling experts").

250. *Report of the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 27.

State judges have adopted new frameworks for considering admissibility that incorporate new scientific developments: through interpretation of state constitutional provisions, through evidence law, by tasking state commissions with revising jury instructions, and through the judges' supervisory power. A series of courts have concluded, for example, "as a matter of state constitutional law, that it is appropriate to modify [due process standards] to conform to recent developments in social science and the law."[251] Courts increasingly recognize that vision is limited by uncertainty and bias, where "research further reveals that an array of variables can affect and dilute memory and lead to misidentifications."[252] State courts have adopted new jury instructions for eyewitness evidence, new standards for in-court identifications, and new admissibility standards, as well as requiring that pretrial reliability hearings (validity hearings) be conducted.[253] As in federal courts, most state courts now permit, within the discretion of the trial judge, the appointment of expert witnesses on eyewitness perception and memory and may find it to be an error to exclude an eyewitness expert that would play an important role in educating the jury.[254] Further, state lawmakers have been quite active in this area, enacting a series of statutes that regulate eyewitness identification procedures, and also, to a lesser extent, how judges review such evidence when procedures do not comply with these statutes.[255]

The sections that follow describe: (1) the U.S. Supreme Court's constitutional rulings regarding eyewitness evidence; (2) the rulings by federal courts of appeals and lower federal courts; (3) state court rulings, including those creating new frameworks for reviewing eyewitness evidence, and those regarding expert evidence and jury instructions; and (4) state statutes regarding eyewitness evidence.

The U.S. Supreme Court

The U.S. Supreme Court first examined the constitutional criminal procedure implications of eyewitness identifications in a trilogy of cases decided in 1967. In *Gilbert v. California*[256] and *United States v. Wade*,[257] the Court held that, once indicted, a person has a right under the Sixth Amendment to have a defense lawyer present at an in-person live lineup.[258] Further, in *Wade*, the Court held that a failure to provide defense counsel at a postindictment lineup does not

251. State v. Harris, 191 A.3d 119, 134 (Conn. 2018).

252. State v. Henderson, 27 A.3d 872, 247 (N.J. 2011).

253. *See* Thomas D. Albright & Brandon L. Garrett, *The Law and Science of Eyewitness Evidence*, 102 Boston U. L. Rev. 511, 591 & Appendix B (2022) (surveying state court rulings).

254. *See id.* at 593 & Appendix C (surveying leading state cases).

255. *See id.* at 580 & Appendix A (surveying state statutes).

256. Gilbert v. California, 388 U.S. 263 (1967).

257. United States v. Wade, 388 U.S. 218, 228 (1967).

258. *Id.* at 236–38; *Gilbert*, 388 U.S. at 272.

result in suppression of the evidence if the witness had an "independent source" for the identification, based on factors including the opportunity to observe the culprit at the crime scene, any discrepancies in witness descriptions, any prior identifications or failure to identify the person, and the lapse of time between the act and the lineup.[259] The Court later held, in its 1973 ruling in *United States v. Ash*, that right to counsel does not extend to photo array procedures, which, today, police use far more than live or in-person lineups.[260]

In *Stovall v. Denno*,[261] the third case in the 1967 trilogy, the Supreme Court held that the Due Process Clause also regulates eyewitness identification procedures, stating that certain procedures may be "so unnecessarily suggestive" that the identification evidence must be suppressed.[262] However, the Court rejected any per se rule against the use of showup identifications and held that, based on a review of the totality of the circumstances, such identifications could be admissible despite the risks of suggestion.[263] Further, in its 1968 ruling in *Simmons v. United States*,[264] the Court emphasized the overall reliability of the eyewitness's identification in finding no grounds for suppressing the identification under the due process theory announced in *Stovall*.[265]

In 1972, in *Neil v. Biggers*,[266] the Court found, in a case involving a trial that occurred pre-*Stovall*, that even if an identification was conducted in an unnecessarily suggestive manner, a court should consider five factors in examining whether there was a "likelihood of misidentification" and a due process remedy warranted:

> the opportunity of the witness to view the criminal at the time of the crime, the witness' degree of attention, the accuracy of the witness' prior description of the criminal, the level of certainty demonstrated by the witness at the confrontation, and the length of time between the crime and the confrontation.[267]

In 1977, in *Manson v. Brathwaite*, the Supreme Court adopted this *Biggers* test for all cases, as a general due process rule for consideration of eyewitness identification evidence potentially affected by undue law enforcement suggestion.[268] The Court emphasized that "reliability is the linchpin in determining the admissibility of identification testimony."[269] Although the Court's due process rule

259. *Wade*, 388 U.S. at 241.
260. United States v. Ash, 413 U.S. 300, 321 (1973).
261. Stovall v. Denno, 388 U.S. 293 (1967).
262. *Id.* at 301–02.
263. *Id.*
264. Simmons v. United States, 390 U.S. 377 (1968).
265. *Id.* at 378, 384–86.
266. Neil v. Biggers, 409 U.S. 188 (1972).
267. *Id.* at 199–200.
268. Manson v. Brathwaite, 432 U.S. 98, 114 (1977) (holding *Biggers* factors should be used to assess reliability).
269. *Id.*

asks whether police used suggestive identification procedures, any such suggestiveness can be excused based on a review of the five "reliability" factors set out in *Biggers*.[270] The Court did not assign any particular weight to these factors.[271]

In the years since the ruling, scientific research has advanced considerably, and in ways that might inform how a court applies the test. One central concern that scientists have raised concerning this test is that it could permit judges to excuse suggestive eyewitness identification procedures, based on an assumption that other indicia of reliability might counteract the influence of police suggestion.[272] The NRC report emphasized:

> [T]he test treats factors such as the confidence of a witness as independent markers of reliability when, in fact, it is now well established that confidence judgments may vary over time and can be powerfully swayed by many factors.[273]

However, in applying the test, the *Biggers* factors need not be considered as "independent," and judges have discretion regarding what weight to attach to them. Thus, the application of the *Manson* test could be informed by more recent scientific research.

Subsequent rulings by the Supreme Court have not altered the test adopted in *Manson,* but they have added further discretion and also limitations to the inquiry. In *Watkins v. Sowders*, the Supreme Court held that an evidentiary hearing regarding identification testimony, outside the presence of the jury, need not be conducted in all cases, but remains within the discretion of the trial judge.[274] The Court noted that it may often be advisable to do so outside the presence of the jury but that in some situations it would not be needed, and cross-examination of the eyewitness can preserve a defendant's due process rights.[275]

Most recently, the Supreme Court held in *Perry v. New Hampshire* that when eyewitness misidentifications are not due to intentional police action, the Due Process Clause does not apply.[276] In that case, the suspect was detained near the

270. *Id.*

271. *Id.* at 116 (stating only that *Biggers* factors are "for the jury to weigh").

272. Wells & Quinlivan, *supra* note 20, at 16 (condemning federal courts' application of *Manson* test to find that identifications were reliable even when surrounding procedures were "highly suggestive").

273. *See* Identifying the Culprit, *supra* note 12.

274. Watkins v. Sowders, 449 U.S. 341, 349 (1981).

275. *Id.* For examples of lower federal courts emphasizing that cross-examination sufficiently addressed reliability concerns, see, e.g., *United States v. Maloney*, 513 F.3d 350, 356 (3d Cir. 2008); *United States v. Saint Louis*, 889 F.3d 145, 154–55 (4th Cir. 2018); *Amador v. Quarterman*, 458 F.3d 397, 415 (5th Cir. 2006); *United States v. Stokes*, 631 F.3d 802, 807 (6th Cir. 2011).

276. Perry v. New Hampshire, 565 U.S. 228, 245 (2012) ("The fallibility of eyewitness evidence does not, without the taint of improper state conduct, warrant a due process rule requiring a trial court to screen such evidence for reliability before allowing the jury to assess its creditworthiness.").

crime scene and the eyewitness looked out of her apartment, saw him there, and made an identification.[277] The majority did note that they "do not doubt either the importance or the fallibility of eyewitness identifications" but maintain that state evidence law and safeguards such as expert testimony and jury instructions should be relied on to ensure accurate presentation of evidence.[278]

Lower Federal Court Rulings

In 1972, in *United States v. Telfaire*, the Court of Appeals for the District of Columbia Circuit crafted particularly influential jury instructions regarding eyewitness evidence.[279] The instructions explain that jurors "must consider the credibility of each identification witness in the same way as any other witness, consider whether he is truthful, and consider whether he had the capacity and opportunity to make a reliable observation on the matter covered in his testimony." The instructions ask jurors to: examine the "circumstances under which the identification was made"; "scrutinize the identification with great care"; "consider the length of time that lapsed between the occurrence of the crime and the next opportunity of the witness to see the defendant"; and consider any instances in which the witness failed to identify the defendant.

The instructions are fairly brief and generically touch on certain factors that can be relevant to assessing the reliability of eyewitness evidence. The *Telfaire* ruling was a step forward at the time, encouraging judges to provide additional context regarding eyewitness evidence. To be sure, the ruling predates the body of modern research on the subject, and as we will describe, courts have more recently offered instructions explicitly grounded in scientific research. While most federal courts use the *Telfaire* instructions,[280] other federal courts adopt a "flexible approach" that provides district courts with discretion whether to use eyewitness jury instructions should they conclude, based on "strong reliability" using the *Biggers/Manson* factors, that no such instruction is necessary.[281] Thus, those federal courts rely on the factors set out in *Manson v. Brathwaite* in deciding whether to give the jury the *Telfaire* jury instructions.[282] Some federal courts

277. *Id.*

278. *Id.*

279. United States v. Telfaire, 469 F.2d 552 (D.C. Cir. 1972) (per curiam).

280. United States v. Anderson, 739 F.2d 1254, 1258 (7th Cir. 1984); *see also* United States v. Greene, 591 F.2d 471, 475 (8th Cir. 1979).

281. *See, e.g.*, United States v. Luis, 835 F.2d 37, 41 (2d Cir. 1987) ("We believe this flexible approach remains the better course because it avoids imposing rigid requirements on trial courts under the threat that failure to give the requested charge will later be grounds for automatic reversal.").

282. *Telfaire*, 469 F.2d at 558–59 (providing model jury special instructions on identification).

have also modestly supplemented the *Telfaire* instructions to include additional factors, such as whether it was a cross-racial identification.[283]

On the question of admissibility of expert evidence regarding eyewitness memory, federal courts, like state courts, have in recent years become more receptive to the use of such experts.[284] Thus, in 1999, the Eighth Circuit explained it was "especially hesitant" to find it an abuse of discretion not to admit eyewitness expert testimony unless the case rests "exclusively on uncorroborated eyewitness testimony."[285] Other courts, however, emphasized the usefulness of eyewitness expert testimony and found it at times an abuse of discretion not to permit the defense such an expert.[286] Pre-*Daubert,* the Eleventh Circuit was the only federal court of appeals that had adopted a per se rule that eyewitness expert evidence would *not* be permitted.[287] The Eleventh Circuit still maintains that approach, while every other court of appeals now permits such experts, subject to the discretion of the district judge.[288]

More recent federal rulings regarding experts on eyewitness evidence have relied more heavily on modern scientific research to explain the need to provide such research to the jury. For example, the Second Circuit, in its 2021 ruling in *United States v. Nolan*, found defense counsel per se ineffective for failing to call an expert witness on eyewitness evidence. The panel explained: "*Strickland* ordinarily does not require defense counsel to call any particular witness," but given

283. Comm. on Model Crim. Jury Instructions for the Third Circuit, Model Criminal Jury Instructions § 4.15 (2017).

284. For an overview, see Lauren Tallent, *Through the Lens of Federal Evidence Rule 403: An Examination of Eyewitness Identification Expert Testimony Admissibility in the Federal Circuit Courts*, 68 Wash. & Lee L. Rev. 765, 777–78 (2011).

285. *See* United States v. Villiard, 186 F.3d 893, 895 (8th Cir. 1999).

286. *See, e.g.*, United States v. Rodriguez–Felix, 450 F.3d 1117, 1125 (10th Cir. 2006) ("[A]n expert's testimony describing how certain factors, falling outside a typical juror's experience, may affect an eyewitness's identification is the very type of scientific knowledge to which *Daubert*'s relevance prong is addressed."); United States v. Stevens, 935 F.2d 1380, 1400 (3d Cir. 1991) (reversing where misidentification was the key issue at trial and expert's proposed testimony was outside the realm of typical juror knowledge); United States v. Moore, 786 F.2d 1308, 1313 (5th Cir. 1986) ("In some cases casual eyewitness testimony may make the entire difference between a finding of guilt or innocence. In such a case expert eyewitness identification testimony may be critical."); United States v. Smith, 736 F.2d 1103, 1106 (6th Cir. 1984) (recognizing that expert testimony on eyewitness identification "might have refuted [jurors'] otherwise common assumptions about the reliability of eyewitness identification"); *see also* Bell v. Miller, 500 F.3d 149, 155–56 (2d Cir. 2007); United States v. Brownlee, 454 F.3d 131, 144 (3d Cir. 2006); Ferensic v. Birkett, 501 F.3d 469, 477–80 (6th Cir. 2007).

287. United States v. Holloway, 971 F.2d 675, 679 (11th Cir. 1992).

288. United States v. Owens, 682 F.3d 1358, 1359 (11th Cir. 2012) (memorandum) (Barkett, J., dissenting from denial rehearing en banc) ("Having been given the opportunity to change our court's position that appellate courts are never permitted to review for abuse of discretion the exclusion of expert testimony regarding the reliability of eyewitness identifications, we should avail ourselves of it. That isolated position, established thirty years ago, conflicts with all of the other circuits and all but five of the states that have considered the question.").

the centrality of the eyewitness identification evidence, the failure to call an expert was a prejudicial "fail[ure] to render reasonable professional assistance." The Sixth Circuit explained, in a 2007 ruling finding exclusion of the defense eyewitness expert to be unconstitutional, that the significance of an expert's testimony "cannot be overstated" because, without it, a jury has "no basis beyond defense counsel's word to suspect the inherent unreliability" of an eyewitness identification.[289]

The Third Circuit Court of Appeals convened a task force on eyewitness identifications that issued a detailed report in 2019 that surveyed scientific research regarding eyewitness evidence.[290] The report was not binding, but rather made recommendations to law enforcement, advocates, and courts, including recommendations regarding conducting lineups, videotaping lineups, documenting witness confidence, blind administration, as well as continuing education materials for lawyers. The task force did not make recommendations concerning amending jury instructions, noting that a separate Third Circuit committee considers model jury instructions for the court. Those instructions already include recommendations for a range of additions to the *Telfaire* instructions.[291]

State Court Rulings

Several state courts have modified the federal due process rule, relying on the research that has developed in the intervening decades since *Manson*. Several state courts have adopted different factors in a "refinement" of the *Manson* and *Biggers* factors.[292] Additional state courts, however, have rejected the *Manson* test, substituting a new framework based on scientific research.[293] Still additional states have adopted changes to jury instructions and expert evidence regarding eyewitnesses.

289. *Ferensic,* 501 F.3d at 477–80.

290. *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 27.

291. Comm. on Model Crim. Jury Instructions for the Third Circuit, Model Criminal Jury Instructions § 4.15 (2017), *supra* note 283.

292. *See, e.g.*, State v. Ramirez, 817 P.2d 774, 780–81 (Utah 1991) (altering three "reliability" factors to focus on effects of suggestion), *abrogated by* State v. Lujan, 459 P.3d 992 (2020); State v. Marquez, 967 A.2d 56, 69–71 (Conn. 2009) (adopting detailed criteria for assessing suggestion); Brodes v. State, 614 S.E.2d 766, 771 & n.8 (Ga. 2005) (rejecting use of eyewitness certainty); State v. Hunt, 69 P.3d 571, 576 (Kan. 2003) (adopting five factor "refinement" of federal due process test); State v. Almaraz, 301 P.3d 242, 253 (Idaho 2013) ("[B]y outlining the system and estimator variables that research has convincingly shown to impact the reliability of eye-witness identification, we hope to provide guidance to lower courts applying the test. . . .").

293. State v. Henderson, 27 A.3d 872, 877 (N.J. 2011) (finding scientific evidence shows *Manson* test should be revised); State v. Lawson, 291 P.3d 673, 678 (Or. 2012) (en banc) (revising *Manson* test in light of scientific evidence); Commonwealth v. Gomes, 22 N.E.3d 897, 909 (Mass. 2015) (finding scholarly research should inform identification instructions); Young v. State, 374 P.3d 395,

The most notable rulings have been by the Supreme Courts of Alaska (*Young v. Alaska*, 2016), Connecticut (*State v. Guilbert*, 2012), Hawaii (*State v. Cabagbag*, 2012), New Jersey (*State v. Henderson*, 2011), Oregon (*State v. Lawson*, 2012), Utah (*State v. Clopten*, 2009), and Wisconsin (*State v. Dubose*, 2005).[294]

The first prominent judicial ruling involving a wholesale reconsideration of the federal due process framework was the New Jersey Supreme Court ruling in *State v. Henderson*. The court-appointed special master "evaluate[d] scientific and other evidence about eyewitness identifications[,] . . . presided over a hearing that probed testimony by seven experts and produced more than 2,000 pages of transcripts along with hundreds of scientific studies," and then issued a detailed report.[295] The decision set out a framework, grounded in scientific research, in which pretrial hearings must examine eyewitness identification evidence to assess its reliability.[296] A defendant must show evidence of unreliability, and the state must counter with evidence of reliability. In response, the judge may consider remedies, including jury instructions at trial.

In 2016, the Alaska Supreme Court revised the *Manson* test and adopted a more detailed framework, asking judges to focus on a range of variables that affect the reliability of eyewitness evidence, as in the *Henderson* decision, and adjudicating these questions pretrial, including with the benefit of expert testimony to explain how those variables may apply to different factual situations.[297] If the defendant meets this burden, "the trial court should suppress the evidence—both the pretrial identification and any subsequent in-court identification by the witness." If the defendant does not, however, then "the court should admit the evidence and provide the jury with an instruction appropriate to the context of the case." The approach, thus, resembles the type of functional framework adopted in New Jersey. In 2018, in *State v. Harris,* the Connecticut Supreme Court adopted a framework much like that in New Jersey (noting the overlap with that approach), holding expert testimony may be valuable, and "it may be appropriate for the trial court to craft jury instructions to assist the jury in its consideration of this issue."[298]

427 (Alaska 2016) (urging trial courts to incorporate evolving scientific understanding to supplement test variables).

294. *See, e.g.*, *Young*, 374 P.3d at 427; State v. Guilbert, 49 A.3d 705 (Conn. 2012); State v. Cabagbag, 277 P.3d 1027, 1035–38 (Haw. 2012) (holding that when identification is central issue, court must, at defendant's request, give jury instruction about factors affecting reliability); *Henderson*, 27 A.3d 872; State v. Lawson, 291 P.3d 673 (Or. 2012); State v. Clopten, 223 P.3d 1103 (Utah 2009); State v. Dubose, 699 N.W.2d 582 (Wis. 2005). The *Cabagbag* ruling prompted model jury instructions in Hawaii. State v. Kaneaiakala, 450 P.3d 761, 774 (Haw. 2019) (holding that when applicable, jury must receive instructions to consider impact of suggestive procedures on identification reliability).

295. *Kaneaiakala*, 450 P.3d at 877.

296. *Henderson*, 27 A.3d. at 894–96, 925–26 (noting research shows "that memory is a constructive, dynamic, and selective process").

297. *Young*, 374 P.3d at 413.

298. *Harris*, 191 A.3d at 134.

Adopting a slightly different approach, the Hawaii Supreme Court ruled in 2019 that "courts must, at minimum, consider any relevant factors set out in the Hawaii Standard Instructions governing eyewitness and show-up identifications, as may be amended." Thus, the court did not set out a series of factors, as the New Jersey court did, but rather set out jury instructions that may be changed over time.[299] Tracking the New Jersey focus on system variables, the court also emphasized the importance of suggestion: "[T]rial courts must also consider the effect of the suggestiveness on the reliability of the identification in determining whether it should be admitted into evidence."

In 2015, the Massachusetts Supreme Judicial Court "review[ed] the scholarly research, analyses by other courts, amici submissions," and the report by a Massachusetts Supreme Judicial Court Study Group on Eyewitness Identification.[300] The court recommended judges provide a set of more concise jury instructions on eyewitness identification evidence. The approach similarly includes a range of research-informed factors, but adopts jury instructions that are more manageable in length.

In contrast to these approaches, which rely on a multifactored standard for admissibility and instructions for educating jurors, the Oregon Supreme Court has endorsed review of reliability of eyewitness evidence relying on a general Rule 403 analysis under the Oregon Rules of Evidence. The Oregon court identified two problems with the prior framework: The "threshold requirement of suggestiveness inhibits courts from considering evidentiary concerns" and the "inquiry fails to account for the influence of suggestion on evidence of reliability."[301]

Other state courts have made more selective changes, such as rejecting eyewitness confidence as a factor to be considered when assessing the validity of an eyewitness identification,[302] or requiring additional jury instructions.[303]

299. *Kaneaiakala*, 450 P.3d at 777.

300. Commonwealth v. Gomes, 22 N.E.3d 897, 900 (Mass. 2015).

301. State v. Lawson, 291 P.3d 673, 748 (Or. 2012).

302. State v. Harris, 191 A.3d 119, 134 (Conn. 2018). 191 A.3d at 134 (expanding protections for suggestiveness of eyewitness identification procedure); State v. Almaraz, 301 P.3d 242, 253 (Idaho 2013) (naming estimator and system variables to be considered in applying *Manson* test); People v. Adams, 423 N.E.2d 379, 383–84 (N.Y. 1981) (concluding where witness makes identification under influence of suggestive procedure and there is no independent source that suggests defendant is perpetrator, court should exclude evidence); State v. Ramirez, 817 P.2d 774, 781 (Utah 1991) (modifying *Manson* test to remove level of certainty as factor), *abrogated by* State v. Lujan, 459 P.3d 992 (2020); State v. Hunt, 69 P.3d 571, 572 (Kan. 2003) (refining Neil v. Biggers, 409 U.S. 188 (1972), approach by adding factors from *Ramirez*); State v. Discola, 184 A.3d 1177 (Vt. 2018) (rejecting witness certainty as factor in assessing reliability).

303. State v. Kaneaiakala, 450 P.3d 761, 777–78 (Haw. 2019) (holding when identification has been procured through suggestive procedure or when central to case, jury must be instructed to consider potential impact of suggestive procedures on reliability of identification).

In addition, several courts have limited the use of courtroom identifications.[304] The Massachusetts Supreme Judicial Court and the Connecticut Supreme Court have ruled that no in-court identification is permitted if an out-of-court identification was suppressed as unduly suggestive.[305] The Massachusetts Supreme Judicial Court has also ruled that first-time courtroom identifications should not normally be permitted.[306] Other courts adopt a burden-shifting approach toward in-court identifications.[307]

State Eyewitness Evidence Statutes

Lawmakers in almost half of the states have required law enforcement adoption of best practices for eyewitness evidence. State legislation has required the study of eyewitness procedures, development of model policies, or increasingly, outright adoption of eyewitness identification practices by law enforcement. To date, twenty-five states have enacted legislation requiring law enforcement adoption of written eyewitness identification procedures.[308] These statutes were often enacted to ensure uniformity in adoption of best practices. Most of those

304. For a proposal that prior identifications should be admitted, but in-court identifications and statements of confidence should not, see Brandon L. Garrett, *Judging Eyewitness Evidence*, 104 Judicature 30, 34–35 (2020), https://perma.cc/5ETW-QUV5.

305. *See* Commonwealth v. Johnson, 45 N.E.3d 83, 92–94 (Mass. 2016) (reasoning that "a subsequent in-court identification cannot be more reliable than the earlier out-of-court identification, given the inherent suggestiveness of in-court identifications and the passage of time"); State v. Dickson, 141 A.3d 810, 817 n.2 & n.3 (Conn. 2016) ("[I]n cases in which identity is an issue, in-court identifications that are not preceded by a successful identification in a nonsuggestive identification procedure implicate due process principles and, therefore, must be prescreened by the trial court." (footnotes omitted)).

306. *See* Commonwealth v. Crayton, 21 N.E.3d 157, 169 (Mass. 2014) ("Where an eyewitness has not participated before trial in an identification procedure, we shall treat the in-court identification as an in-court showup, and shall admit it in evidence only where there is 'good reason' for its admission.").

307. *See* State v. Hickman, 330 P.3d 551, 568 (Or. 2014), *modified on reconsideration*, 343 P.3d 634 (2015) ("Courts considering the admissibility of first-time in-court identifications generally have placed the burden of seeking a prophylactic remedy on the defendant."); United States v. Domina, 784 F.2d 1361, 1369 (9th Cir. 1986), *cert. denied*, 479 U.S. 1038 (1987) (noting that district court's denial of request for in-court lineup will only be overturned if in-court identification procedures were so suggestive and conducive to irreparable misidentification as to amount to denial of due process).

308. *See* Cal. Penal Code § 859.7 (West 2022); Colo. Rev. Stat. § 16-1-109 (2022); Conn. Gen. Stat. § 54-1p (2021); Fla. Stat. § 92.70 (2021); Ga. Code Ann. § 17-20-2 (2021); Hawaii Rev. Stat. 801k (2022); 725 Ill. Comp. Stat. 5/107A-2 (2021); Kan. Stat. Ann. § 22-4619 (2021); La. Code Crim. Proc. Ann. Arts. 251–53 (2021); Md. Code Ann., Pub. Safety § 3-506 (West 2021); Minn. Stat. § 626.8433 (2021); Neb. Rev. Stat. § 81-1455 (2021); Nev. Rev. Stat. § 171.1237 (2021); N.H. Rev. Stat. Ann. § 595-C:2 (2022); N.M. Stat. Ann. § 29-3B-3 (2022); N.Y. Crim. Proc. Law § 60.25 (Mckinney 2021); N.C. Gen. Stat. § 15A-284.52 (2021); Ohio Rev. Code Ann. § 2933.83 (West 2021); Okla. Stat. Tit. 22, § 21 (2021); Tex. Code Crim. Proc. Ann. Art. 38.20 (West 2021); Utah Code Ann. § 77-8-4 (West 2021); Vt. Stat. Ann. Tit. 13, § 5581 (2022); Va. Code Ann. § 19.2-390.02 (2021); W. Va. Code §§ 62-1E-1 to -2 (2021); Wis. Stat. § 175.50 (2022).

states regulate the eyewitness identification procedures to be used, including blinded lineups (see above section titled "Communication during lineups: The importance of blinded lineup administration"), clear written instructions, and documenting the confidence of an eyewitness.[309] In addition, seventeen states have revised their jury instructions, departing from the *Telfaire* model or models reciting factors from *Manson v. Brathwaite*.[310]

For example, Massachusetts developed model jury instructions regarding eyewitness identifications that include both preliminary language, to be read prior to opening statements, and final instructions.[311] The preliminary language advises the jury that "[t]he mind does not work like a video recorder" and cautions that factors during and after the observed event can alter an individual's memory of that event. The final instructions are far more detailed, describing identification as a three-stage process: perception of an event, storage of information about the event in one's mind, and later recollection of that stored information.[312] The instructions caution that a variety of factors, each of which is listed in the instructions themselves, may affect accuracy of a memory at any stage of the process. In selecting the factors to be included in its revised jury instructions, Massachusetts relied on "a near consensus [standard] in the relevant scientific community."[313] The analysis suggested five such principles (although without referring to factors that affect the validity of visual perceptual experience):

1. human memory does not operate like a video recording that a person can replay to recall what happened;
2. a witness's level of confidence in an identification may not indicate its accuracy;
3. high levels of stress can reduce the likelihood of making an accurate identification;
4. information from other witnesses or outside sources can affect the reliability of an identification and inflate an eyewitness's confidence in the identification; and

309. *See* Albright & Garrett, *supra* note 253, at 511, Appendix A.

310. Those states are Alaska, Connecticut, Florida, Georgia, Hawaii, Kansas, Maine, Maryland, Massachusetts, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Utah, and Virginia.

311. Mass. Sup. Jud. Ct., Model Jury Instructions on Eyewitness Identification: Preliminary/Contemporaneous Instruction 2 (2015); Mass. Sup. Jud. Ct., Model Jury Instructions on Eyewitness Identification: Model Eyewitness Identification Instruction 2 (2015) (establishing jury instruction to be given where jury heard evidence affirmatively identifying defendant and such identification is contested).

312. Model Jury Instructions on Eyewitness Identification: Model Eyewitness Identification Instruction 2 (Mass. Sup. Jud. Ct. 2015).

313. *Id.* at 2–10.

5. viewing the same person in multiple identification procedures may increase the risk of misidentification.[314]

We note that there is not strong evidence that these jury instructions fully accomplish their aim to better educate jurors regarding the potential strengths and weaknesses of eyewitness evidence, including because jurors place such powerful weight on courtroom identifications and courtroom expressions of confidence.[315] Indeed, research has not found that the instructions adopted in New Jersey, or the more concise Massachusetts instructions, are effective in helping jurors distinguish between high- and low-confidence initial identifications.[316] More recent research suggests that instructions explaining *why* it is important to place weight on initial confidence, and not courtroom identifications and courtroom confidence, may help jurors make that important distinction.[317]

Police Practice Recommendations

In recent years, professional policing organizations, such as the IACP, Major Cities Chiefs Association, and the CALEA, have taken active roles in promoting best practices, including through model policies.[318] The Third Circuit task force emphasized that results of research on eyewitness procedures must be translated to law enforcement personnel with proper guidelines for maximum effectiveness. The U.S. Department of Justice adopted a policy in 2017, applying to all federal law enforcement agencies, that requires blind administration of lineups, among other procedures.[319] Twenty-six states and the federal government have voluntarily adopted such policies regulating eyewitness identifications.[320] The model policies generally require blinded lineups (see above section titled "Communication during lineups: The importance of blinded lineup administration"),

314. Ralph D. Gants & Erik N. Doughty, *Where Science Conflicts with Common Sense: Eyewitness Identification Reform in Massachusetts*, 79 Alb. L. Rev. 1617, 1625–26 (2017) (quoting Commonwealth v. Gomes, 22 N.E.3d 897, 903, 909 (Mass. 2015)).

315. *See generally,* Garrett et al., *supra* note 33.

316. *Id.*

317. Garrett et al., *supra* note 34.

318. *See* U.S. Dep't of Just./Int'l Ass'n of Chiefs of Police, National Summit on Wrongful Convictions: Building A Systemic Approach to Prevent Wrongful Convictions 13–14 (2013), https://perma.cc/ML3T-2ZJE; Comm'n on Accreditation for L. Enf 't, CALEA L. Enf't Agency Standards § 42.2.11 (2010); Int'l Ass'n of Chiefs of Police Pol'y Ctr., Model Pol'y: Eyewitness Identification 1–4 (2016).

319. Procedures for Conducting Photo Arrays, *supra* note 7.

320. The states are Arkansas, Colorado, Connecticut, Delaware, Florida, Georgia, Kansas, Kentucky, Louisiana, Maine, Maryland, Michigan, Minnesota, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, Rhode Island, Utah, Vermont, Virginia, Washington, and Wisconsin.

clear written instructions, and documenting confidence; many recommend videotaping. Policing organizations have incorporated research recommendations into lineup policies and training; federal support has been important to such efforts.[321]

National organizations have also set out recommendations for law enforcement regarding incorporation of research into eyewitness identification procedures. The American Law Institute, as part of its Principles of Policing project, adopted principles for law enforcement concerning eyewitness identifications.[322] In 2020, the AP-LS released a scientific review paper summarizing research in the field and making new recommendations.[323]

Conclusion

The evolving law of eyewitness evidence shows how science can play a pivotal role in our legal system. While constitutional precedent can change slowly, its practical application can be informed by scientific research. Evidentiary practices, like appointment of experts, pretrial hearings and rulings, and jury instructions, can also be informed by scientific research. We have experienced a paradigm shift: Almost every jurisdiction—through judicial rulings, legislation, model policy, and police policy—has embraced scientific research regarding eyewitness evidence. These approaches aim to more accurately collect evidence from eyewitnesses, and to prevent contamination of that evidence through suggestion and other forms of bias.

321. These efforts began in the late 1990s with the National Institute of Justice's Technical Working Group for Eyewitness Evidence. *See* U.S. Dep't of Just., Nat'l Inst. of Just., Eyewitness Evidence: A Guide For Law Enforcement, *supra* note 27 (putting forth recommendations to improve procedures for collection and preservation of eyewitness evidence within criminal justice system).

322. Principles of the Law: Policing ch. 10 (A.L.I. 2022).

323. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3.

Glossary of Terms

accuracy. The probability that the identification is of the perpetrator. Estimated by predictive variables in real casework and commonly measured in known source tests as the number of correct identifications relative to the total number of identifications (correct + incorrect).

attentional hijacking. The condition in which a highly salient stimulus directs the focus of attention away from an intended target of attention.

bias. To evaluate something in a prejudicial manner. The meaning here is not pejorative: It refers to prior knowledge or experience that is used to infer the meaning of an incomplete or ambiguous sensory event.

binary classification. A cognitive process in which the individual must decide between two discrete alternatives, as in the lineup participant is a match or a nonmatch to memory of the perpetrator of a crime.

blinded test. (Also known as "single-blinded test.") A classification task in which the witness is prevented (hence "blinded") from knowing the status of lineup participants. See double-blinded test.

categorical decision. A decision in favor of one of a set of discrete alternatives, such as classification of a lineup participant as a match or a non-match.

categorical perception. The experience in which different sensory stimuli of the same category are perceived similarly, which entails loss of sensitivity to individual sensory differences within a category.

center of gaze. The region of visual space corresponding to the portion of the retina with the highest density of photoreceptors; the region with the highest visual acuity.

color blindness. A genetically determined condition in which an individual has fewer than three types of color-specific photoreceptors, which limits the range of discriminable colors.

confidence. A cognitive state and expression that reflects perceived certainty of a decision, as in perceived certainty that one's identification of the perpetrator is correct.

confidence inflation. The increase in perceived certainty in a decision that follows from reinforcement and perceived corroboration by others.

contextual variables. Variables associated with a signal event, as in environmental, perceptual, or cognitive states that are correlated with the experience of witnessing a crime and identifying a perpetrator. These variables may be useful in assessing the accuracy of the identification.

contrast sensitivity. A quantified measure of an observer's sensitivity to a difference between two sensory stimuli, such as different intensities of light, which bears on the ability to detect and discriminate patterns.

cross-race effect. A well-established perceptual phenomenon in which the ability of an observer to distinguish between faces of members of their own race exceeds the ability to distinguish between faces of members of a different race. The effect is believed to be a consequence of perceptual learning resulting from predominant exposure of individuals to same-race faces during critical stages of development. It has the obvious potential to yield differential error rates in same-race versus cross-race eyewitness identifications.

decision criterion. The threshold value of a decision variable that an observer uses to distinguish two different classifications of a sensory stimulus, as in the specific degree of perceptual similarity between a lineup face and a face stored in memory that distinguishes between a match versus a nonmatch decision.

decision variable. The measured or computed variable that an observer uses to make a classification decision, as in the similarity between the appearance of a lineup face and the eyewitness's memory of the face of the culprit.

declarative memory. Long-term memory for facts and events that can be retrieved at will and brought into conscious awareness.

degree of visual angle. The angular extent of visual space from the nodal point of the eye, which is commonly used to refer to the spatial extent of an object as optically projected onto the back of the eye. In the context of object recognition, it is used to determine the ability to resolve features in the projected image.

diagnosticity ratio (DR). The ratio of the probability of correct identification of a known culprit to the probability of incorrect identification of a known innocent suspect. Used in laboratory studies of eyewitness performance as a measure of the ability of a witness to distinguish the culprit from an innocent person. Not a true measure of discriminability, since the diagnosticity ratio is also influenced by the decision criterion employed by the eyewitness.

discriminability. Reflects the degree to which an instrument of measurement can distinguish the values of two inputs, assessed independently of the decision criterion employed to make the classification decision. In the eyewitness context, this refers to the extent to which the witness can tell the difference between the culprit and an innocent suspect.

double-blinded test. A classification task in which the witness and the administrator of the task are both "blinded" from knowing the status of lineup participants; this is modeled after an element of the scientific method designed to prevent decision bias based on prior knowledge of experimental conditions.

ecological validity. The extent to which variables, results, and conclusions from an experimental study generalize to the real world.

episodic memory. The form of declarative memory that holds information about the sequence or narrative content of memory. Eyewitness memories are of this type.

estimator variable. Contextual factors associated with the viewing conditions and perceptual/cognitive state of the witness at the time of the crime. These include environmental factors such as intensity of illumination and duration of events, as well as observer variables, such as acuity, emotion, and pain. These variables are not controllable in an eyewitness context. However, because they can predictably impair the fidelity of visual perception and memory encoding, they thus have great potential to inform the trier of fact as to the probability that an identification is accurate.

face blindness. Laboratory studies have shown that face recognition ability varies continuously across the human population. At one extreme of this distribution are "face-blind" people who are extremely poor at recognition memory for faces. The degree of face blindness can be quantified by now-standard face recognition tests. Performance on such tests is an estimator variable that can be used for quantitative prediction of the accuracy of eyewitness reports. Contrasted with face super-recognizers.

face super-recognizer. Laboratory studies have shown that face recognition ability varies continuously across the human population. At the extreme opposite end of the continuum from face-blind people are face super-recognizers, who excel at recognition memory for faces. The degree of performance can be quantified by now-standard face recognition tests, which can serve as quantitative predictors of the accuracy of eyewitness reports.

fair lineup. A lineup composed of fillers who all bear a similar perceptual relationship to the witness's reported description of the culprit. Fair lineups are less likely to elicit biased identifications and wrongful convictions of people who stand out from the crowd based on similarity to the witness's description.

gaze direction. The point in visual space that is the momentary target of the center of gaze, which possesses the highest visual acuity.

illusory conjunctions. A perceptual phenomenon in which briefly viewed attributes of a scene become disjoined from the objects they originate from and conjoined with others, as in a red letter D and a blue number 2 that are perceived as a blue D and a red 2.

inattentional blindness. Failure to perceive an otherwise prominent visual stimulus that results from lack of attentional focus directed to the stimulus.

lineup type. The specific type of visual presentation and behavioral task used to test an eyewitness's ability to identify the culprit.

long-term memory. Memories that have been consolidated into long-term storage, where they are less subject to forgetting but still modifiable by the acquisition of new information.

luminance contrast. A measure of the contrasting amounts of light reflected (i.e., luminance) from two spatially adjacent features of a visual scene, which is a primary determinant of an observer's ability to perceptually distinguish the features.

memory consolidation. The transfer of perceptual or cognitive information from labile short-term storage into more stable long-term storage, which is a time-dependent process mediated by changes in neuronal connectivity.

memory encoding. The transfer of perceptual or cognitive experiences into a form storable by the brain, which underlies the ability to learn from those experiences.

memory plasticity. Memory is modifiable over time, to a high degree during acquisition and to a lesser degree once consolidated. Memory systems are more plastic during critical stages of human development, but retain some plasticity throughout life.

memory retention interval. The amount of time that passes between initial acquisition of a new memory and a subsequent test of the fidelity of that memory. Within that retention window, memories predictably fade because of forgetting and can be modified by intrusion of new information.

memory retrieval. The process of bringing a stored piece of information back to a state where it can be used to guide decisions and behavior. Retrieval is commonly elicited by an external sensory stimulus or by an internal association.

memory storage. The state in which memories are maintained for the long term and accessible for retrieval at a later time. Memories are less subject to forgetting during this period but can still be modified by acquisition of new information.

misdirection. Direction of the focus of visual attention away from the location of a behaviorally relevant object or feature. A prominent tool used in magical tricks that demonstrates the easy fallibility of visual perceptual experiences and the inappropriate assertion of confidence in such experiences. This phenomenon underlies the "weapon focus" effect that impairs accuracy of eyewitness identification.

myopia. A visual pathology in which the optics of the eye are misaligned, such that the projected pattern of light on the back surface of the eye is not well focused. Although correctable with external lenses or corneal surgery, this pathology is common enough in the human population, to varying degrees, that it could easily limit the accuracy of eyewitness identification.

noise. Random or irrelevant elements that interfere with detection of coherent and informative signals. In human sensory systems, this interference is frequent and can take the form of distortion of signals at the sensory periphery, such as optical impairment, or by diffraction of light by fluid in the eye. Other sources

of noise include occluding elements of a visual scene or attentional distractions. The result is reduced fidelity of sensation and perceptual uncertainty.

object surface reflectance. The amplitude and chromatic spectrum of white light reflected from the surface of an object. This reflectance is a physical property of the surface itself, which interacts with the amplitude and chromatic spectrum of light illuminating the scene.

operating characteristic. A set of parameters that define the performance of an instrument based on empirical assessment and knowledge of its underlying mechanism. Generally, these describe performance in terms of the validity and reliability of the output for a given type of input. Quantitative performance descriptions of this sort offer a useful systems approach to understanding the abilities and weaknesses of eyewitness reports.

perception. The process by which evanescent sensations are linked to environmental cause and made enduring and coherent through the assignment of meaning, utility, and value.

perceptual improvisation. Unconscious search through memory for information (probabilities of the accuracy of previously acquired knowledge, beliefs, and hypotheses) that enables rapid identification, recognition, and spatial understanding of novel and otherwise unrecognizable stimuli.

perceptual learning. Long-term experience-dependent changes in the way that people perceive sensory stimuli, which can enhance or distort perceptual experience. Apropos of eyewitness identification, the cross-race effect, which comes about through extended selective exposure to faces of people of one's own race, is a consequence of perceptual learning.

priors. Shorthand for "prior probability," which refers to the probability of accuracy for knowledge, beliefs, or hypotheses about the world that an observer holds prior to a signal event, such as witnessing a crime. Used synonymously with prejudices or biases, priors enable an observer to unconsciously fill in blanks in the presence of perceptual uncertainty.

receiver operating characteristic (ROC). A standard modeling approach to the evaluation of performance on a classification task, drawn from signal detection theory. In the eyewitness context, the ROC is generally presented as a curve that plots the probability of correct detection of a target (the culprit) against the probability of incorrect detection of a non-target (an innocent suspect). The area under the ROC curve is a single-value measure of the ability to discriminate the culprit from an innocent suspect, which has the advantage of being independent of the decision criterion used by the eyewitness. Use of this approach has transformed analysis of data from laboratory studies of factors that affect eyewitness performance, in part because it yields a true measure of discriminability.

recognition memory. The ability to determine whether a present sensory stimulus (a "cue") is the same as a stimulus previously seen and memorized, and subsequently retrieved from memory. In the eyewitness context, a positive outcome from this similarity assessment is the event that leads to an assertion of identification.

reliability. A property of an instrument of measurement (e.g., an eyewitness) that reflects the degree to which it yields the same result every time it is used to measure the same thing. Often used inappropriately to refer to the validity of an instrument of measurement.

retina. The sheet of neuronal tissue that lines the back of the eye. Light reflected from surfaces in the visual environment passes through the lens of the eye and is projected onto the retina, which includes the photoreceptors that are responsible for phototransduction (conversion of luminous energy into neuronal energy). Numerous optical and neuronal noise sources can create visual uncertainty at this stage of processing, undermining the accuracy of perceptual experience.

semantic memory. The form of declarative memory that holds information about meanings of experiences and acquired knowledge of the world.

sensation. The consequence of spatially and temporally patterned stimulation of biological sensory receptors. Antecedent to and contrasted with perception, through which sensations are linked to environmental cause.

sequential lineup. A lineup identification procedure in which each face in the lineup is presented to the witness in isolation, and the complete set is presented in a temporal sequence. Believed to promote absolute lineup comparisons.

short-term memory (working memory). The memory system that has limited capacity for volume of content and duration of storage, but where content is actively maintained and readily available for use to guide decisions and actions (i.e., it is "working" memory). Short-term memory is highly labile storage, subject to distortion or loss in the presence of distractions. Extended maintenance of content requires consolidation into long-term memory storage.

signal detection. From signal detection theory, which is a theoretical framework for understanding the processes and consequences of decision-making in the presence of uncertainty. Used for decades as an efficient means to analyze behavioral responses to sensory stimulation.

simultaneous lineup. A lineup identification procedure in which all faces in the lineup are presented to the witness at the same time.

source memory failure. Loss of the ability to remember how one acquired a specific memory, while the memory itself is intact. A common consequence of this failure is that the source of the memory is wrongly attributed based on priors and may incorrectly modify other memories. This is a memory

system analogue of the introduction of biases to resolve uncertainty in visual perception.

system variables. Procedures and states that may influence the accuracy of an eyewitness report and are potentially under the control of the criminal justice system, since they bear on activities that occur after the witnessed events. Includes lineup type and blinded versus nonblinded lineup administration.

target absent lineup. A type of lineup procedure that does not include the culprit ("target") as one of the lineup participants. Commonly refers to an empirical practice in laboratory studies of eyewitness performance, which is used to assess the probability that witnesses mistakenly identify an innocent suspect. Results (probabilities of incorrect identifications) are compared to those obtained using a target present lineup. The ratio of correct to incorrect identification probabilities is the diagnosticity ratio.

target present lineup. A type of lineup procedure that includes the culprit ("target") as one of the lineup participants. Commonly refers to an empirical practice in laboratory studies of eyewitness performance, which is used to assess the probability that witnesses correctly identify the culprit. Results (probabilities of correct identifications) are compared to those obtained using a target absent lineup. The ratio of correct to incorrect identification probabilities is the diagnosticity ratio.

task irrelevant information. Contextual information associated with a visual or cognitive task that is not required to solve the task but is at risk of biasing the outcome. Apropos of eyewitness identification, task irrelevant information can refer to reports from journalists, family, friends, and legal counsel that may artifactually influence a witness's degree of certainty in their identification. Contrasted with task relevant information, which is needed to solve the task.

uncertainty. In the eyewitness context, the state of being uncertain about a measured piece of sensory information, as a result of noise and ambiguity. The common consequence is that uncertainty is resolved by completion of perceptual experience based on prior knowledge or beliefs.

unfair (biased) lineup. A lineup in which the chosen fillers possess different degrees of similarity of appearance to the witness's description of the culprit. Unfair lineups can bias the choices made by a witness because some faces stand out from the crowd, which may lead to wrongful investigation, prosecution, and conviction.

validity. The property of an instrument of measurement that reflects the degree to which it yields the result that it was intended to measure. In the eyewitness context, this generally refers to the measured probability that eyewitnesses will correctly identify the culprit, which is typically obtained from laboratory studies in which ground truth is known.

visual acuity. A measure of the clarity and sharpness of vision, empirically assessed by the threshold ability to resolve points in an image at varying degrees of angular separation. A measure of an eyewitness's visual acuity is essential to predicting the accuracy of an identification.

visual crowding. Reduction of the probability of correctly detecting and perceiving a target stimulus when it is viewed in a dense arrangement of other stimuli. In the eyewitness context, this is manifested as increasing limits on the ability to correctly perceive the actors and events of a crime as a function of the density of actors and objects in the scene.

visual performance. A generic term for the quantifiable ability of an observer to accurately detect and discriminate visual stimuli.

visual search. A well-studied perceptual phenomenon in which the amount of time it takes to detect a specific object amid a set of related objects is dependent on the presence or absence of distinctive features. In the presence of distinctive features, such as a unique color, immediate perceptual "pop-out" occurs. If, on the other hand, the object is defined by a unique conjunction of features (for example, a single red letter T in a field of green Ts and both red and green Ls), time to target object detection is predictably delayed because the visual system must "search" serially through the scene until it is encountered. This attentional delay has the potential to markedly increase uncertainty in time-limited eyewitness events.

Reference Guide on Statistics and Research Methods

DAVID H. KAYE AND HAL S. STERN

David H. Kaye, M.A., J.D., is Regents Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, and Distinguished Professor of Law and Academy Professor Emeritus, Pennsylvania State University School of Law.

Hal S. Stern, Ph.D., is Distinguished Professor, Department of Statistics, University of California, Irvine.

Authors' Note: Research for this reference guide was completed in 2023. Professor David A. Freedman co-authored the first three editions of this reference guide. His writing and thinking are evident throughout this edition as well.

CONTENTS

FIGURES

TABLES

Introduction

Statistical assessments are prominent in many kinds of legal cases, including antitrust, employment discrimination, toxic torts, and voting rights cases. This reference guide describes the elements of statistical reasoning. We hope the explanations will help judges and lawyers to understand statistical terminology, to see the strengths and weaknesses of statistical arguments, and to apply relevant legal doctrine. The guide is organized as follows:

- This introduction provides an overview of the field, discusses the admissibility of statistical studies, and offers some suggestions about procedures that encourage the best use of statistical evidence.
- The section titled "How Have the Data Been Collected?" addresses data collection. It explains why the design of a study is the most important determinant of its quality. The section compares experiments with observational studies and surveys with censuses, indicating when the various kinds of study are likely to provide useful results.
- The section titled "How Have the Data Been Presented?" discusses the art of summarizing data. This section considers the mean, median, and standard deviation. These are basic descriptive statistics, and most statistical analyses use them as building blocks. This section also discusses patterns in data that are brought out by graphs, percentages, and tables.
- The section titled "What Inferences Can Be Drawn from the Data?" describes the logic of statistical inference, emphasizing foundations and disclosing limitations. This section covers estimation, standard errors and confidence intervals, *p*-values, and hypothesis tests.
- The section titled "Correlation and Regression" shows how associations can be described by scatter diagrams, correlation coefficients, and regression lines. Regression is often used in attempts to infer causation from association. This section explains the technique, indicating the circumstances under which it and other statistical models are likely to succeed—or fail.
- Advances in computing speed and the availability of large datasets have made it possible to fit models of greater complexity than those described in the section titled "Correlation and Regression." The section titled "Data Science and Statistical Machine Learning" describes the developing fields of "data science" and "machine learning" and statistical issues that affect the confidence one can have in the output of machine-learning techniques.
- An appendix discusses the scope of the theory of probability, conditional probabilities, Bayes' rule, and perspectives on statistical inference.
- The glossary defines statistical terms that may be encountered in litigation, including ones that do not appear in the body of the guide.

Admissibility and Weight of Statistical Studies

Statistical studies suitably designed to address a material issue generally will be admissible under the Federal Rules of Evidence. The hearsay rule rarely is a serious barrier to the presentation of statistical studies, because such studies may be offered to explain the basis for an expert's opinion or may be admissible under the learned treatise exception to the hearsay rule.[1] Most statistical methods applied in litigation are described in textbooks or journal articles and are capable of producing useful results when properly applied. As such, these methods generally satisfy important aspects of the "scientific knowledge" requirement in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[2] However, a particular study may use a method that is entirely appropriate but so poorly executed that it should be inadmissible under Federal Rules of Evidence 403 and 702.[3] Or, the method may be inappropriate for the problem at hand and thus lack the "fit" referred to in *Daubert*.[4] Or the study might rest on data of the type not reasonably relied on by statisticians or substantive experts and hence not be suitable as a basis for an opinion under Rule 703. Often, however, the battle over statistical evidence concerns weight or sufficiency rather than admissibility.

Varieties and Limits of Statistical Expertise

For convenience, the field of statistics may be divided into three subfields: probability theory, theoretical statistics, and applied statistics. Probability theory is the mathematical study of outcomes that are governed, at least in part, by chance. Theoretical statistics is about understanding the properties of statistical procedures, including error rates; probability theory plays a key role in this endeavor. Applied statistics draws on both these fields to develop techniques for collecting or analyzing particular types of data.

Statistical expertise is not confined to witnesses with degrees in statistics. Because statistical reasoning underlies many kinds of empirical research, scholars in a variety of fields—including biology, business, economics, epidemiology, medicine, political science, psychology, and sociology—are exposed to statistical ideas, with an emphasis on the methods most important to the discipline. The diffusion of statistical concepts and methods across so many fields raises the question

1. *See generally* 2 McCormick on Evidence §§ 321, 324.3 (Robert P. Mosteller ed., 8th ed. 2020). Studies published by government agencies also may be admissible as public records. *Id.* § 296.

2. 509 U.S. 579, 589–90 (1993).

3. *See* Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999) (suggesting that the trial court should "make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field").

4. *Daubert*, 509 U.S. at 591.

of who is qualified to conduct and testify to statistical assessments—a statistical methodologist, a substantive scientist, or both? Much depends on context.

If the study involves assembling and then analyzing case-specific data, the choice of which data to examine and how best to model a particular process could require both subject-matter and general statistical expertise. The two types of expertise might be combined in a single individual. A labor economist, for example, should be able to supply a definition of the relevant labor market from which an employer draws its employees. This economist also might be sufficiently educated in statistical tests for group differences to compare the race of new hires to the racial composition of the labor market. When the analysis is as straightforward as the comparison of two proportions, a single substantive expert may suffice.[5] But further analysis of the possible reasons for the racial disparity might require input from an expert with an understanding of more advanced methods. This expert might be a statistician or an econometrician, or a scientist or engineer who works with other data in a completely different field of study.

The critical question is whether the individual has the education and experience to determine which statistical techniques are appropriate for the task at hand and to apply such techniques with an appreciation of their limitations. Experts who specialize in using statistical methods, and whose professional careers demonstrate this orientation, are most likely to use appropriate procedures and correctly interpret the results. To ascertain the extent to which an expert has this orientation and acumen, one can look to formal education, professional accomplishments, and reputation in the pertinent community of experts. Has the individual studied quantitative methods? Taught them? Used them in research? Which ones? If the expert is or was an academic professional, a university teaching portfolio with courses on statistical methods is a good sign. Ideally, the publication and research record will include studies that use or describe the same or similar methods as those applied (or applicable) to the case at bar.[6] Invitations from mainstream journal editors to review articles submitted for publication, from academic conference organizers to present research involving statistics, and from government regulatory and research agencies to serve on expert panels that evaluate statistical work are further indications of recognized expertise within a statistical discipline. General membership in most scientific and professional societies requires no special accomplishments, but certain designations, prizes, or awards from such organizations for scholarship are a sign that an

5. Of course, the case could be presented in the form of interlocking testimony from two experts—the labor economist followed by any expert with the appropriate expertise in the method for making the statistical comparison. The substantive knowledge may be of lesser value in selecting a statistical model. Various models might be consistent with the substantive knowledge, and there may be purely statistical criteria for choosing among them.

6. Academic experts are likely to have a long list of publications, but length alone is not the best indication of pertinent expertise. Not all publishers are equal, and every field has a relatively small number of top-tier journals.

expert's work has had an impact. Of course, these factors are merely indicia of degrees of expertise and specialization. Highly competent work can be done by individuals with shorter curricula vitae.

Again, not every case involving computations of probability or statistics necessitates a highly skilled statistical specialist. Medical practitioners and forensic scientists who are not statistical specialists often make statistical assessments of evidence or rely on and present the results of statistical studies. In these situations, it is important to ensure that the witnesses do not exceed the bounds of their statistical expertise. The scientist or practitioner might lack basic information about the studies underlying their testimony. *State v. Garrison*[7] illustrates the problem. In this murder prosecution involving bitemark evidence, a dentist was allowed to testify that "the probability factor of two sets of teeth being identical in a case similar to this is, approximately, eight in one million," even though "he was unaware of the formula utilized to arrive at that figure other than that it was 'computerized.'"[8] Likewise, laboratory analysts or examiners may have only a limited understanding of the foundations of automated systems for comparing patterns within DNA samples, toolmarks, fingerprints, voice recordings, and the like. Once the formulas or methods have been shown to work well, and when their limitations and proper use are known, the practitioners may be qualified to operate the statistical machinery, as it were, without an in-depth understanding of the routinized or automated statistical or probabilistic method. But the operational training may not qualify them to explain the developmental research. They may have to limit their statistical testimony to an explanation of how they arrived at their estimates and to statements of the extent to which software or hardware is used in practice and whether there are publications on its validity.

Procedures That Enhance Statistical Testimony

Maintaining Professional Autonomy

Ideally, experts who conduct research in the context of litigation should proceed with the same objectivity that would be required in other professional contexts. Thus, experts who testify (or who supply results used in testimony) should conduct the analysis required to address, in a professionally responsible fashion, the issues posed by the litigation.[9] Questions about the freedom of inquiry accorded

7. 585 P.2d 563 (Ariz. 1978).

8. *Id.* at 566, 568. For other examples, see David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12.2 (2d ed. 2011).

9. *See* Nat'l Research Council Panel, The Evolving Role of Statistical Assessments as Evidence in the Courts 164 (Stephen E. Fienberg ed., 1989) [hereinafter NRC Panel] (recommending that the expert be free to consult with colleagues who have not been retained by any party to the litigation and that the expert receive a letter of engagement providing for these and other safeguards).

to testifying experts, as well as the scope and depth of their investigations, may reveal some of the limitations to the testimony.

Disclosing Limitations and Other Analyses

Statisticians analyze data using a variety of methods. To permit a fair evaluation of the analysis that is eventually settled on, the testifying expert can be asked how that approach was developed, whether alternative approaches were considered, and, if so, what the results were.[10] Ethical guidelines for statisticians require them to disclose known and suspected limitations in the data and its analysis.[11]

Disclosing Data and Analytical Methods Before Trial

The collection of data often is expensive and subject to errors and omissions. Moreover, careful exploration of the data can be time-consuming. To minimize debates at trial over the accuracy of data and the choice of analytical techniques, pretrial-discovery procedures should be used, particularly with respect to the quality of the data and the method of analysis.[12] In some cases, these could include requirements for specifying in detail the study design and analysis that are planned *before* data collection begins.[13]

10. *Id.* at 167; *cf.* Edith Beerdsen, *Litigation Science After the Knowledge Crisis*, 106 Cornell L. Rev. 529 (2021) (discussing responses to the "replication crisis" in the academic biomedical and behavioral science publication process and suggesting pretrial procedures to make the exercise of "analytical flexibility" in "litigation science" more transparent); Yoav Benjamini, *Selective Inference: The Silent Killer of Replicability*, 2 Harv. Data Sci. Rev. (2020), https://doi.org/10.1162/99608f92.fc62b261 (noting the problem of presenting or emphasizing selected findings within scientific publications).

11. ASA Comm. on Professional Ethics, Ethical Guidelines for Statistical Practice, Feb. 2022, at 3 & 4, https://perma.cc/P4P6-JU6C. Invoking an attorney's professional duty not to intentionally mislead the courts on the facts or the law, the Panel on Statistical Assessments as Evidence insisted that "it is not appropriate for the attorney to seek to avoid such revelation [of alternative forms of data and analyses] by consulting a series of experts without revealing to the experts ultimately retained any prior history of the involvement of other experts in the litigation." NRC Panel, *supra* note 9, at 167.

12. *See* The Special Comm. on Empirical Data in Legal Decision Making, Recommendations on Pretrial Proceedings in Cases with Voluminous Data, *reprinted in* NRC Panel, *supra* note 9, app. F; David H. Kaye, *Improving Legal Statistics*, 24 Law & Soc'y Rev. 1255 (1990).

13. Drawing on "open science" practices, commentators have proposed adaptations of practices for "registration" of academic research studies. *See* Beerdsen, *supra* note 10; section titled "Recent Emphasis on the Limitations of *p*-values" below.

How Have the Data Been Collected?

The interpretation of data often depends on understanding "study design"—the plan for a statistical study and its implementation.[14] Different designs are suited to answering different questions. Also, flaws in the data can undermine any statistical analysis, and data quality is often determined by study design.

In many cases, statistical studies are used to show causation. Do food additives cause cancer? Does capital punishment deter crime? Would additional disclosures in a securities prospectus cause investors to behave differently? The design of studies to investigate causation is the first topic of this section.[15]

Sample data can be used to describe a population. The population is the whole class of units that are of interest; the sample is the set of units chosen for detailed study. Inferences from the part to the whole are justified when the sample is representative. Sampling is the second topic of this section.

Finally, issues associated with the reliability and accuracy of collected data will be considered. Measurement error should be assessed and the likely impact of errors considered. Data quality is the third topic of this section.

All the sections concern the study of "variables." In statistics, a variable is a characteristic of the units in a study. With a study of people, the unit of analysis is the person, and the variables describe people. Two such variables would be income (dollars per year) and educational level (years of schooling completed). With a study of school districts, the unit of analysis is the district. Typical variables include average family income of district residents and average test scores of students in the district.

Variables may be related to one another in various ways. Many studies examine whether a variable or group of variables, known as independent variables, are related to an outcome or dependent variable. For example, census data can be analyzed to determine whether people who complete more years of school tend to have higher incomes later in life. Educational level would be the independent variable, and annual income the dependent variable. In a study of smoking and lung cancer, the independent variable could be smoking (perhaps measured by the number of cigarettes smoked per day), and the dependent variable could mark the presence or absence of lung cancer.

14. For introductory treatments of data collection, see, for example, David Freedman et al., *Statistics* (4th ed. 2007); Darrell Huff, How to Lie with Statistics (1993); David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (10th ed. 2019); Hans Zeisel, *Say It with Figures* (6th ed. 1985); Hans Zeisel & David Kaye, *Prove It with Figures: Empirical Methods in Law and Litigation* (1997).

15. *See also* Steve C. Gold et al., *Reference Guide on Epidemiology,* "The Different Kinds of Epidemiologic Studies" section, in this manual.

Is the Study Designed to Investigate Causation?

Types of Studies

When causation is the issue, anecdotal evidence can be brought to bear. So can observational studies or controlled experiments. Anecdotal reports may be of value, but they are ordinarily more helpful in generating lines of inquiry than in proving causation. In medicine, evidence from clinical practice can be a good starting point for discovery of cause-and-effect relationships, but the anecdotal experience of practitioners is not definitive.[16] Observational studies can establish that one factor is associated with another, but work is needed to bridge the gap between association and causation. Randomized controlled experiments are ideally suited for demonstrating causation.

Anecdotal evidence usually amounts to reports that events of one kind are followed by events of another kind. Typically, the reports are not even sufficient to show association, because there is no comparison group. For example, some children who live near power lines develop leukemia. Does exposure to electrical and magnetic fields cause this disease? The anecdotal evidence is not compelling because leukemia also occurs among children without exposure.[17] It is necessary to compare disease rates among those who are exposed and those who are not. If exposure causes the disease, the rate should be higher among the exposed and lower among the unexposed. That would be association.

16. Consequently, many courts have suggested that attempts to infer causation from anecdotal reports are inadmissible as unsound methodology under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993). *See, e.g.,* Miller v. Pfizer, Inc., 356 F.3d 1326, 1331 (10th Cir. 2004) (affirming the district court's exclusion in part because "placing substantial emphasis on a few challenge-dechallenge-rechallenge studies and case reports is not a generally accepted methodology"); Hendrix *ex rel.* G.P. v. Evenflo Co., Inc., 609 F.3d 1183 (11th Cir. 2010) ("Case studies and clinical experience, used alone and not merely to bolster other evidence, are . . . insufficient to show general causation."); *cf.* Matrixx Initiatives, Inc. v. Siracusano, 563 U.S. 27, 44 (2011) (concluding that adverse-event reports combined with other information could be of concern to a reasonable investor and therefore subject to a requirement of disclosure under SEC Rule 10b-5, but stating that "the mere existence of reports of adverse events . . . says nothing in and of itself about whether the drug is causing the adverse events"). Other courts are more open to "differential diagnoses" based primarily on timing. *E.g.*, Best v. Lowe's Home Ctrs., Inc., 563 F.3d 171 (6th Cir. 2009) (reversing the exclusion of a physician's opinion that exposure to propenyl chloride caused a man to lose his sense of smell because of the timing in this one case and the physician's inability to attribute the change to anything else).

17. *See* Nat'l Research Council, Possible Health Effects of Exposure to Residential Electric and Magnetic Fields (1997). There are problems in measuring exposure to electromagnetic fields, and results are inconsistent from one study to another. For such reasons, the epidemiologic evidence for an effect on health is inconclusive. Nat'l Cancer Inst., Electromagnetic Fields and Cancer, May 30, 2022 ("Studies have examined associations of these cancers with living near power lines, with magnetic fields in the home, and with exposure of parents to high levels of magnetic fields in the workplace. No consistent evidence for an association between any source of non-ionizing EMF and cancer has been found.").

The next issue is crucial: Exposed and unexposed people may differ in ways other than the exposure they have experienced. For example, children who live near power lines could come from poorer families and be more at risk from other environmental hazards. Such differences can create the appearance of a cause-and-effect relationship. Other differences can mask a real relationship. Cause-and-effect relationships often are quite subtle, and carefully designed studies are needed to draw valid conclusions.

An epidemiological classic makes the point. At one time, it was thought that lung cancer was caused by fumes from tarring the roads, because many lung cancer patients lived near roads that recently had been tarred. This is anecdotal evidence. But the argument is incomplete. For one thing, most people—whether exposed to asphalt fumes or unexposed—did not develop lung cancer. A comparison of rates was needed. Epidemiologists found that exposed persons and unexposed persons suffered from lung cancer at similar rates: Tar was probably not the causal agent. Exposure to cigarette smoke, however, turned out to be strongly associated with lung cancer. This study, in combination with later ones, made a compelling case that smoking cigarettes is the main cause of lung cancer.[18]

A good study design compares outcomes for subjects who are exposed to some factor (the treatment group) with outcomes for other subjects who are not exposed (the control group). With comparison groups, there is another important distinction—between controlled experiments and observational studies. In a controlled experiment, the investigators decide which subjects will be exposed and which subjects will be in the control group. In observational studies, the researchers do not determine which subjects are exposed; often, the subjects themselves choose their exposures. Because of self-selection or for other reasons, the treatment and control groups are likely to differ with respect to influential factors other than the ones of primary interest. These other factors are called lurking variables or confounding variables.[19] A confounding variable can be correlated with the independent variable and act causally on the dependent variable. When

18. Richard Doll & A. Bradford Hill, *A Study of the Aetiology of Carcinoma of the Lung*, 2 Brit. Med. J. 1271 (1952), https://doi.org/10.1136/bmj.2.4797.1271. This was a matched case-control study. Cohort studies soon followed. *See* Gold et al., *supra* note 15. For a review of the evidence on causation, see 38 Int'l Agency for Research on Cancer (IARC), World Health Org., IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans: Tobacco Smoking (1986) (updated 100E, A Review of Human Carcinogens: Personal Habits and Indoor Combustions 43–211 (2012)).

19. Epidemiologists sometimes limit "confounding" to "a bias due to the existence of a common cause of exposure and outcome" and define "selection bias" as "bias by [selecting units for study based] on common effects of otherwise unrelated variables." Stephen R. Cole et al., *Illustrating Bias Due to Conditioning on a Collider*, 39 Int'l J. Epidemiology 417, 420 (2010), https://doi.org/10.1093/ije/dyp334. The distinction can be helpful in spotting different threats to causal inference, but this section uses the terms more loosely.

that happens, the confounder—not the independent variable—could be responsible for differences seen on the dependent variable. With the health effects of power lines, family background is a possible confounder; so is exposure to other hazards. Many confounders have been proposed to explain the association between smoking and lung cancer, but careful epidemiological studies have ruled them out, one after the other.

Confounding remains a problem even for the best observational research. For example, women with herpes are more likely to develop cervical cancer than other women. Some investigators concluded that herpes caused cancer: In other words, they thought the association was causal. Later research showed that the primary cause of cervical cancer was human papilloma virus (HPV). Herpes was a marker of sexual activity. Women who had multiple sexual partners were more likely to be exposed not only to herpes but also to HPV. The association between herpes and cervical cancer was due to other variables.[20]

Randomized Controlled Experiments

In randomized controlled experiments, investigators assign subjects to treatment or control groups at random. The groups are therefore likely to be comparable, except for the treatment. This minimizes the role of confounding. Minor imbalances will remain, owing to the play of random chance; the likely effect on study results can be assessed by statistical techniques.[21] The bottom line is that causal inferences based on well-executed randomized experiments are generally more secure than inferences based on well-executed observational studies.

The following example should help bring the discussion together. Today, we know that taking aspirin helps prevent heart attacks. But initially, there was some controversy. People who take aspirin rarely have heart attacks. This is anecdotal evidence for a protective effect, but it proves almost nothing. After all, few people have frequent heart attacks, whether or not they take aspirin regularly. A good study compares heart-attack rates for two groups: people who take aspirin (the treatment group) and people who do not (the controls). An observational study would be easy to do, but in such a study the aspirin-takers are likely to be different from the controls. Indeed, they are likely to be sicker—that is why they are taking aspirin. The study would be biased against finding a protective effect.

20. For additional examples and further discussion, see David A. Freedman, *From Association to Causation: Some Remarks on the History of Statistics,* 14 Stat. Sci. 243 (1999). Some studies find that herpes is a "cofactor," which increases risk among women who are also exposed to HPV. Only certain strains of HPV are carcinogenic.

21. Randomization of subjects to treatment or control groups puts statistical tests of significance on a secure footing. *See* section titled "What Inferences Can Be Drawn from the Data?" below.

Randomized experiments are harder to do, but they provide better evidence.[22] The experiments demonstrate the protective effect.[23]

In summary, data from a treatment group without a control group generally reveal very little and can be misleading. Comparisons are essential. If subjects are assigned to treatment and control groups at random, a difference in the outcomes between the two groups can usually be accepted, within the limits of statistical error,[24] as a good measure of the treatment effect. However, if the groups are created in any other way, differences that existed before treatment may contribute to differences in the outcomes or mask differences that otherwise would become manifest. Observational studies succeed to the extent that the treatment and control groups are comparable—apart from the treatment.

Observational Studies

The bulk of the statistical studies seen in court are observational, not experimental. Take the question of whether capital punishment deters murder. To conduct a randomized controlled experiment, people would need to be assigned randomly to a treatment group or a control group. People in the treatment group would know they were subject to the death penalty for murder; the controls would know that they were exempt. Conducting such an experiment is not possible.

Many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention. Researchers have catalogued differences in the incidence of murder in states with and without the death penalty and have analyzed changes in homicide rates and execution rates over the years. When reporting on such observational studies, investigators may speak of "control groups" (e.g., the states without capital punishment) or claim they are "controlling for" confounding variables by statistical methods such as multiple regression.[25] However, association is not causation. The

22. One important feature of a clinical trial is "blinding" to prevent unconscious bias. In a "double blind" trial, neither the patients nor the clinicians ascertaining the outcomes know who received the treatment and who did not. This prevents the expectations of the subjects and the experimenters from systematically affecting the observed outcomes in one group relative to the other.

23. But randomized experiments also show that aspirin can cause internal bleeding, raising the practical questions of whether and when a low-dose regimen is more beneficial than harmful. *See, e.g.,* U.S. Preventive Services Task Force, Aspirin Use to Prevent Cardiovascular Disease: Preventive Medication, Apr. 26, 2022, https://perma.cc/C8V9-WMYS. In other instances, experiments have contradicted strongly held beliefs. *E.g.*, Eric A. Klein et al., *Vitamin E and the Risk of Prostate Cancer: Results of the Selenium and Vitamin E Cancer Prevention Trial (SELECT)*, 306 JAMA 1549 (2011), https://doi.org/10.1001/jama.2011.1437.

24. *See* section titled "What Inferences Can Be Drawn from the Data?" below.

25. Multiple regression is described in Daniel L. Rubinfeld & David Card, *Reference Guide on Multiple Regression and Advanced Statistical Models*, in this manual. On the limits of regression to cope

causal inferences that can be drawn from analysis of observational data—no matter how complex the statistical technique—usually rest on a foundation that is less secure than that provided by randomized controlled experiments.

When an external change in circumstances naturally creates a treatment and a control group in a manner that is comparable to random assignment by the researchers, the study may be called a "natural experiment" or a "quasi-experiment."[26] A celebrated example comes from Dr. John Snow's investigation of cholera and sewage in the mid-1800s. The residents of an area in London received water from two companies with overlapping water mains drawing from a part of the Thames River that was heavily polluted with raw sewage. Then, one company moved its water intake upstream, to a less polluted spot. Snow showed that during a London cholera epidemic soon afterwards, the death rates from cholera in homes with the less polluted water was less than one-eighth that for the more polluted homes—and less than that for the rest of London.[27] Snow argued that "no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this . . . ."[28] In modern terminology, the argument is that even though the study was observational, the *seemingly* random assignment of homes to the two intake points protects against confounding and bias.

Observational studies can be very useful even when assignments to the groups being compared do not resemble randomization imposed by an experimenter. For example, there is strong observational evidence that smoking causes lung cancer (see section titled "Types of Studies" above). Generally, observational studies provide good evidence in the following circumstances:

- The association is seen in studies with different designs, on different kinds of subjects, and done by different research groups.[29] That reduces the chance that the association is due to a defect in one type of study, a peculiarity in one group of subjects, or the idiosyncrasies of one research group.

with lurking variables, see Richard A. Berk, *Regression Analysis: A Constructive Critique* (2004); Richard Berk, *What You Can and Can't Properly Do with Regression*, 26 J. Quantitative Criminology 481 (2010), https://doi.org/10.1007/s10940-010-9116-4; David A. Freedman, *Statistical Models: Theory and Practice* (2005).

26. *See, e.g.,* Frank de Vocht et al., *Conceptualising Natural and Quasi Experiments in Public Health*, 21 BMC Med. Research Methodology 32 (2021), https://doi.org/10.1186/s12874-021-01224-x.

27. John Snow, On the Mode of Transmission of Cholera 55–98 (2d ed. 1855).

28. *Id*. at 46.

29. For example, case-control studies are designed one way and cohort studies another, with many variations. *See, e.g.,* David D. Celentano & Moyses Szklo, *Gordis Epidemiology* (6th ed. 2020); *supra* note 18.

- The association holds when effects of confounding variables are taken into account by appropriate methods, for example, comparing smaller groups that are relatively homogeneous with respect to the confounders.[30]
- There is a plausible explanation for the effect of the independent variable; alternative explanations in terms of confounding should be less plausible than the proposed causal link.[31] Thus, evidence for the causal link does not depend on observed associations alone.

Observational studies can produce legitimate disagreement among experts, and there is no mechanical procedure for resolving such differences of opinion. In the end, deciding whether associations are causal typically is not a matter of statistics alone, but also rests on scientific judgment.[32]

There are, however, some basic questions to ask when appraising causal inferences based on empirical studies:

- Was there a control group? Unless comparisons can be made, the study has little to say about causation.
- If there was a control group, how were subjects assigned to treatment or control: through a process under the control of the investigator (a controlled experiment) or through a process outside the control of the investigator (an observational study)?
- If the study was a controlled experiment, was the assignment made using a chance mechanism (randomization), or did it depend on the judgment of the investigator?

If the data came from an observational study or a nonrandomized controlled experiment,

- How did the subjects come to be in treatment or in control groups?
- Are the treatment and control groups comparable?

30. The idea is to control for the influence of a confounder by stratification—making comparisons separately within groups for which the confounding variable is nearly constant and therefore has little influence over the variables of primary interest. For example, smokers are more likely to get lung cancer than nonsmokers. Age, gender, social class, and region of residence are all confounders, but controlling for such variables does not materially change the relationship between smoking and cancer rates.

31. A. Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 Proc. Royal Soc'y Med. 295 (1965); Alfred S. Evans, Causation and Disease: A Chronological Journey 187 (1993).

32. At best, statistical analysis can help address the question of how impactful an unknown confounding variable would have to be to vitiate an inference of causation. *See* Jerome Cornfield et al., *Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions*, 22 J. Nat'l Cancer Inst. 173 (1959), https://doi.org/10.1093/jnci/22.1.173; Tyler J. VanderWeele, *Are Greenland, Ioannidis and Poole Opposed to the Cornfield Conditions? A Defence of the E-value*, 51 Int'l J. Epidemiology 364 (2022), https://doi.org/10.1093/ije/dyab218.

- If not, what adjustments were made to address confounding?
- Were the adjustments sensible and sufficient?

Generalizing the Results

In considering what conclusions can be drawn from studies, it is helpful to distinguish between *internal* and *external* validity. Internal validity concerns the specifics of a particular study: Threats to internal validity include confounding and chance differences between treatment and control groups. External validity concerns using a particular study or set of studies to reach more general conclusions. A careful randomized controlled experiment on a large but unrepresentative group of subjects will have high internal validity but low external validity.

Any study must be conducted on certain subjects, at certain times and places, and using certain treatments. To extrapolate from the conditions of a study to more general conditions raises questions of external validity. For example, studies suggest that definitions of insanity given to jurors influence decisions in cases of incest. Would the definitions have a similar effect in cases of murder? Other studies indicate that recidivism rates for ex-convicts are not affected by providing them with temporary financial support after release. Would similar results be obtained if conditions in the labor market were different?

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself. It comes from knowledge about outside factors that would or would not affect the outcome. Such judgments are easiest in the physical and life sciences, but even here, there are problems. For example, it may be difficult to infer human responses to substances that affect animals. First, there are often inconsistencies across test species. A chemical may be carcinogenic in mice but not in rats. Extrapolation from rodents to humans is even more problematic. Second, to get measurable effects in animal experiments, chemicals are administered at very high doses. Results are extrapolated—using mathematical models—to the very low doses of concern in humans. However, there are many dose–response models to use and few grounds for choosing among them. Generally, different models produce radically different estimates of the "virtually safe dose" in humans.[33]

33. David A. Freedman & Hans Zeisel, *From Mouse to Man: The Quantitative Assessment of Cancer Risks,* 3 Stat. Sci. 3 (1988), https://doi.org/10.1214/ss/1177012993; Lorenz R. Rhomberg et al., *Linear Low-Dose Extrapolation for Noncancer Health Effects Is the Exception, Not the Rule*, 41 Critical Reviews Toxicology 1 (2011), https://doi.org/10.3109/10408444.22010.536524. For these reasons, many experts—and some courts in toxic tort cases—have concluded that evidence from animal experiments is generally insufficient by itself to establish causation. Likewise, extrapolation from animals to humans in testing for drug efficacy and safety has been the subject of extensive discussion. Johnique T. Atkins et al., *Pre-clinical Animal Models Are Poor Predictors of Human Toxicities in Phase 1 Oncology Clinical Trials*, 123 Brit. J. Cancer 1496 (2020), https://doi.org/10.1038/s41416-020-01033-x; Brian R. Berridge, *Animal Study Translation: The Other Reproducibility Challenge*,

Sometimes several studies, each having different limitations, all point in the same direction. This combination is why most experts believe that smoking causes lung cancer and many other diseases. So, too, a variety of studies indicate that jurors who approve of the death penalty are more likely to convict in a capital case.[34] Convergent results support the validity of generalizations.

Descriptive Surveys and Censuses

We now turn to a second topic—choosing units for study. A census tries to measure some characteristic of every unit in a population. This is often impractical. Then investigators use sample surveys, which measure characteristics for only part of a population. The accuracy of the information collected in a census or survey depends on how the units are selected for study and how the measurements are made.[35]

What Method Is Used to Select the Units?

By definition, a census seeks to measure some characteristic of every unit in a whole population. It may fall short of this goal; in which case one must ask whether the missing data are likely to differ in some systematic way from the data that are collected.[36] The methodological framework of a scientific survey is different. With probability methods, a sampling frame (i.e., an explicit list of units in the

62 Int'l Lab'y Animal Rsch. J. 1 (2021), https://doi.org/10.1093/ilar/ilac005; John P. A. Ioannidis, *Extrapolating from Animals to Humans*, 12 Sci. Translational Med. 151 (2012), https://doi.org/10.1126/scitranslmed.3004631; Pandora Pound & Merel Ritskes-Hoiting, *Is It Possible to Overcome Issues of External Validity in Preclinical Animal Research? Why Most Animal Models Are Bound to Fail*, 16 J. Translational Med. 304 (2018), https://doi.org/10.1186/s12967-018-1678-1.

34. Phoebe C. Ellsworth, *Some Steps Between Attitudes and Verdicts*, *in* Inside the Juror 42, 46 (Reid Hastie ed., 1993). Nonetheless, in *Lockhart v. McCree*, 476 U.S. 162 (1986), the Supreme Court held that the exclusion of opponents of the death penalty in the guilt phase of a capital trial does not violate the constitutional requirement of an impartial jury.

35. *See* Shari Seidman Diamond et al., *Reference Guide on Survey Research*, "Population Definition and Sampling" section, in this manual.

36. The U.S. Decennial Census does not count everyone that it should, and it counts some people who should not be counted. There is evidence that net undercount is greater in some demographic groups than others. Supplemental studies may enable statisticians to adjust for errors and omissions. *See* Elizabeth Marra & Timothy Kennel, U.S. Census Bureau, PES20-J-01, Source and Accuracy of the 2020 Post-Enumeration Survey Person Estimates: 2020 Post-Enumeration Survey Methodology Report (2022). However, the adjustments rest on uncertain assumptions. *See* Lawrence D. Brown et al., *Statistical Controversies in Census 2000*, 39 Jurimetrics J. 347 (1999); David A. Freedman & Kenneth W. Wachter, *Methods for Census 2000 and Statistical Adjustments, in* Social Science Methodology 232 (Steven Turner & William Outhwaite eds., 2007) (reviewing technical issues and litigation surrounding census adjustment in 1990 and 2000).

population) is created. Individual units then are selected by an objective, well-defined, chance procedure, and measurements are made on the sampled units.

A sampling frame

To illustrate a sampling frame, suppose that a defendant in a criminal case seeks a change of venue. According to the defendant, popular opinion is so adverse that it would be difficult to impanel an unbiased jury. To prove the state of popular opinion, the defendant commissions a survey. The relevant population consists of everyone in the jurisdiction who might be called for jury duty. The sampling frame is the list of all potential jurors, which is maintained by court officials and is made available to the defendant. In this hypothetical case, the fit between the sampling frame and the population would be excellent.

In other situations, the sampling frame is more problematic. In an obscenity case, for example, the defendant can offer a survey of community standards.[37] The population comprises all adults in the legally relevant district, but obtaining a full list of such people may not be possible. Suppose the survey is done by telephone, but cell phones are excluded from the sampling frame. Suppose too that cell phone users, as a group, hold different opinions from landline users. Then the poll is unlikely to reflect the opinions of the cell phone users, no matter how many individuals are sampled and no matter how carefully the interviewing is done.[38]

Selection bias

Many surveys do not use probability methods. In commercial disputes involving trademarks or advertising, the population of all potential purchasers of a product is hard to identify. Pollsters may resort to an easily accessible subgroup of the population—for example, shoppers in a mall. Such convenience samples may be

37. On the admissibility of such polls, see *State v. Midwest Pride IV, Inc.*, 721 N.E.2d 458 (Ohio Ct. App. 1998) (holding one such poll to have been properly excluded, and collecting cases from other jurisdictions); Admissibility of Evidence of Public-Opinion Polls or Surveys in Obscenity Prosecutions on Issue Whether Materials in Question Are Obscene, 59 A.L.R.5th 749.

38. Survey researchers may turn to other methods to reach cell phone users with area codes from the jurisdiction's geographic locale. Kyley McGeeney & Courtney Kennedy, Pew Research Center, Advances in Telephone Survey Sampling (2015), https://perma.cc/D4MY-2L93. However, the cell phone sampling frame will omit people who have moved into the jurisdiction with cell phone numbers from other areas. Again, the mismatch between the sampling frame and the relevant population could bias the results. People who move from county-to-county or state-to-state tend to be younger, more likely to be minority, to be male, and to have lower incomes. *See* Carol Pierannunzi et al., *Sample and Respondent Provided County Comparisons Among Cellular Respondents Using Rate Center Assignments*, 12 Survey Practice 1 (2019), https://doi.org/10.29115/SP-2019-004.

biased by the interviewer's discretion in deciding whom to approach—a form of selection bias—and the refusal of some of those approached to participate—nonresponse bias (see section titled "Of the Units Selected, Which Provide Measurements?" below). Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, interest groups collect information from their members, individuals complete available online surveys, or attorneys choose cases for trial.[39]

A well-known example of selection bias is the 1936 *Literary Digest* poll. After successfully predicting the winner of every U.S. presidential election since 1916, the *Digest* used the replies from 2.4 million respondents to predict that Alf Landon would win the popular vote, 57% to 43%. In fact, Franklin Roosevelt won by a landslide vote of 62% to 38%.[40] The *Digest* was so far off, in part, because it chose names from telephone books, rosters of clubs and associations, city directories, lists of registered voters, and mail order listings.[41] In 1936, when only one household in four had a telephone, the people whose names appeared on such lists tended to be more affluent. Lists that overrepresented the affluent had worked well in earlier elections, when rich and poor voted along similar lines, but the bias in the sampling frame proved fatal when the Great Depression made economics a salient consideration for voters.

There are procedures that attempt to correct for selection bias. In quota sampling, for example, the interviewer is instructed to interview so many women, so many older people, so many ethnic minorities, and the like. But quotas still leave discretion to the interviewers in selecting members of each demographic group and therefore do not solve the problem of selection bias.[42]

Probability methods are designed to avoid selection bias. Once the population is reduced to a sampling frame, the units to be measured are selected by a lottery that gives each unit in the sampling frame a known, nonzero probability of being chosen.[43] Such procedures are used to select individuals for jury duty.

39. *In re* Chevron U.S.A., Inc., 109 F.3d 1016, 1020 (5th Cir. 1997) (although random sampling of 30 cases to resolve common issues or to ascertain damages in 3,000 claims arising from Chevron's allegedly improper disposal of hazardous substances would have been acceptable, having the opposing parties select 15 cases each was not, because those were "not cases calculated to represent the group of 3000 claimants"); *In re* Countrywide Fin. Corp. Mortgage-Backed Sec. Litig., 984 F. Supp. 2d 1021, 1039 (C.D. Cal. 2013) ("The [sample's disproportionate] reliance on loans supporting certificates selected for litigation strikes the Court as a clear example of selection bias . . . ."). *See infra* note 44 (on sampling cases or claims from a large set of similar cases or claims).

40. *See* Freedman et al., *supra* note 14, at 334–35.

41. *Id.* at 335, A-20 n.6.

42. *See id.* at 337–39.

43. Many types of probability sampling have been developed. In simple random sampling, units are drawn at random without replacement. In particular, each unit has the same probability of being chosen for the sample. *Id.* at 339–41. More complicated methods, such as stratified sampling and cluster sampling, give greater selection probabilities to some types of units than others, which has advantages in certain applications. In systematic sampling, every *n*th unit (for example, every

They also have been proposed or used to choose "bellwether" cases for representative trials to resolve issues in a large group of similar cases.[44]

Of the Units Selected, Which Provide Measurements?

Probability sampling ensures that within the limits of chance, the sample will be representative of the sampling frame. But will all these units be measured? When documents are sampled for audit, they can all be examined, at least in principle. Human beings are less easily managed, and some will refuse to cooperate. In the 1936 *Literary Digest* election poll, only 24% of the 10 million people who received questionnaires returned them. Most of the respondents probably had strong views on the candidates and objected to President Roosevelt's economic program. This self-selection is likely to have biased the poll.[45]

Surveys should therefore report nonresponse rates. A large nonresponse rate warns of potential bias.[46] Supplemental studies may establish that nonrespondents are similar to respondents with respect to characteristics of interest, but

fifth, tenth, or hundredth unit) in the sampling frame is selected. If the units are not in any special order, then systematic sampling is often comparable to simple random sampling.

44. *See* Scottsdale Mem'l Health Sys., Inc. v. Maricopa Cnty., 228 P.3d 117, 131–35 (Ariz. Ct. App. 2010) (reviewing major cases and approving in principle of random sampling but concluding that the record failed to demonstrate the adequacy of the statistical sampling plan for resolving 40,000 consolidated cases); David H. Kaye & David A. Freedman, *Statistical Proof*, *in* 1 Modern Scientific Evidence: The Law and Science of Expert Testimony § 5:16, at 398–406 (David L. Faigman et al. eds., 2022–2023) (discussing both legal and statistical issues arising in these cases); David H. Kaye et al., The New Wigmore: A Treatise on Evidence: Expert Evidence § 12.10.3(b) (Cumulative Supp. to 2d ed., 2020) (same).

45. Maurice C. Bryson, *The Literary Digest Poll: Making of a Statistical Myth*, 30 Am. Statistician 184 (1976); Freedman et al., *supra* note 14, at 335–36.

46. For discussions of the admissibility of surveys with extremely low response rates, see *In re ConAgra Foods, Inc.*, 90 F. Supp. 3d 919 (C.D. Cal. 2015) (excluding survey with many defects, including a 95% nonresponse rate); *United States v. H & R Block, Inc.*, 831 F. Supp. 2d 27, 34 (D.D.C. 2011) (98% nonresponse rate does not preclude admission when a judge is the factfinder). On nonresponse rates in studies not undertaken for litigation, see Clearinghouse for Military Family Readiness, Penn. State Univ., *Survey Response Rates: Rapid Literature Review* (2016), available at https://perma.cc/8Y9Z-PTKH.

In *United States v. Gometz*, 730 F.2d 475, 478 (7th Cir. 1984) (en banc), the Seventh Circuit recognized that "a low rate of response to juror questionnaires could lead to the underrepresentation of a group that is entitled to be represented on the qualified jury wheel." Nonetheless, the court held that under the Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), the clerk did not abuse his discretion by failing to take steps to increase a response rate of 30%. According to the court, "Congress wanted to make it possible for all qualified persons to serve on juries, which is different from forcing all qualified persons to be available for jury service." *Gometz*, 730 F.2d at 480. Although it might "be a good thing to follow up on persons who do not respond to a jury questionnaire," the court concluded that Congress "was not concerned with anything so esoteric as nonresponse bias." *Id.* at 479, 482; *cf. In re* United States, 426 F.3d 1 (1st Cir. 2005)

even when demographic characteristics of the sample match those of the population, caution is indicated.[47]

In short, a good survey defines an appropriate population, uses a probability method for selecting the sample, has a high response rate, and gathers accurate information on the sample units. When these goals are met, the sample tends to be representative of the population, and data from the sample can be extrapolated to describe the characteristics of the population. Of course, surveys may be useful even if they fail to meet these criteria. But then, additional arguments are needed to justify the inferences.

Individual Measurements

Is the Measurement Process Reliable?

Reliability and validity are two aspects of accuracy in measurement. In statistics, reliability refers to reproducibility of results. A reliable measuring instrument returns consistent measurements. A scale, for example, is perfectly reliable if it always reports the same weight for the same unchanged object. It may not be accurate—it may always report a weight that is too high or one that is too low—but the perfectly reliable scale always reports the same weight for the same object. Its errors, if any, are systematic: They tend to point in the same direction.

Courts often use "reliable" to mean "that which can be relied on" for some purpose, such as establishing probable cause through a reliable informant or as part of an argument for admitting hearsay statements.[48] Thus, in *Daubert v. Merrell Dow Pharmaceuticals*, the Court distinguished "evidentiary reliability" from reliability in the statistical sense of giving consistent results.[49] This statistical reliability is a component of the broader evidentiary reliability required of scientific evidence. It can be ascertained by measuring the same quantity several times; the measurements must be made independently to avoid bias. Given independence, the correlation coefficient (see section titled "Correlation Coefficients" below) between repeated measurements can be used as a measure of reliability. This is sometimes called a test–retest correlation or a reliability coefficient. But administering the same test twice to the same group of people may be impractical. And

(reaching the same result with respect to underrepresentation of African Americans resulting in part from nonresponse bias).

47. *See* David Streitfeld, *Shere Hite and the Trouble with Numbers*, 1 Chance 26 (1988); Chamont Wang, Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety 174–76 (1992).

48. *E.g.*, United States v. Moore, 824 F.3d 620 (7th Cir. 2016) ("The purpose of Rule 807 is to make sure that reliable, material hearsay evidence is admitted, regardless of whether it fits neatly into one of the exceptions enumerated in the Rules of Evidence."); Fed. R. Evid. 803(18)(B) (to be admissible hearsay, a learned treatise must be a "reliable authority").

49. 509 U.S. 579, 590 n.9 (1993).

even if repeated testing is practical, it may be statistically inadvisable, because subjects may learn something from the first round of testing that affects their scores on the second round. Such "practice effects" are likely to compromise the independence of the two measurements, and independence is needed to estimate reliability. Statisticians therefore use internal evidence from the test itself. For example, a strong correlation between scores on the first half of the test and scores on the second half is evidence of reliability.[50]

The Supreme Court was faced with the imperfect reliability of a test for IQ scores in *Hall v. Florida*.[51] The Court listed various factors contributing to the variability in an individual's test scores, including "health; practice from earlier tests; the environment or location of the test; the examiner's demeanor; the subjective judgment involved in scoring certain questions on the exam; and simple lucky guessing."[52] Having previously held that intellectually disabled offenders could not be punished by execution, the Court held that a state had to account for the potential variability in an individual's IQ score in setting a number above which no one could be deemed intellectually disabled. The particular rule the Court adopted turned on one version of a statistic known as the standard error.[53]

A simpler courtroom example comes from DNA identification. An early method of identification required laboratories to determine the lengths of fragments of DNA. By making independent repeated measurements of the same fragments, laboratories determined the likelihood that two measurements differed by specified amounts.[54] Such results were needed to decide whether a discrepancy between a crime sample and a suspect sample was sufficient to exclude the suspect.[55]

Coding of data also can affect reliability. In many studies, descriptive information is obtained on the subjects. For statistical purposes, the information usually has to be reduced to numbers. The process of reducing information to numbers is called "coding," and the reliability of the process should be evaluated. For example, in a meticulous study of death sentencing in Georgia, legally trained evaluators examined short summaries of cases and ranked them according to the

50. In practice, more refined measures of internal consistency are used to estimate a test's reliability. *See, e.g.,* Neal M. Kingston & Laura B. Kramer, *High Stakes Test Construction and Test Use*, *in* 1 Oxford Handbook of Quantitative Methods: Foundations 189, 201 (Todd D. Little ed., 2013).

51. 572 U.S. 701 (2014).

52. *Id.* at 713.

53. "Standard error" is the subject of the section titled "Is a Difference Statistically Significant?" below. The relationship between the reliability coefficient and different standard errors as well as the choice of a cut-off score that accounts for a given type of standard error is explained in David H. Kaye, *Deadly Statistics: Quantifying an "Unacceptable Risk" in Capital Punishment*, 16 Law, Probability & Risk 7 (2017) (proposing alternatives to the Court's rule).

54. *See* Nat'l Research Council, The Evaluation of Forensic DNA Evidence 139–41 (1996).

55. *Id.*; Nat'l Research Council, DNA Technology in Forensic Science 61–62 (1992). Current methods are discussed in David H. Kaye, *Reference Guide on Human DNA Identification Evidence,* in this manual.

defendant's culpability.[56] Two different aspects of reliability should be considered. First, the "within-observer variability" of judgments should be small—the same evaluator should rate essentially identical cases in similar ways. Second, the "between-observer variability" should be small—different evaluators should rate the same cases in essentially the same way.

Metrologists (specialists in measurement science) similarly distinguish between "reproducibility" and "repeatability." The difference lies in the degree to which the conditions for making measurements are similar. With a repeatable procedure, measurements by the same examiner using the same equipment at the same place and time should be close to one another. With a reproducible procedure, measurements from different examiners even at different places and times also should be similar.[57]

Is the Measurement Process Valid?

Reliability is necessary but not sufficient to ensure accuracy. In addition to reliability, validity is needed. A valid measuring instrument measures what it is supposed to. Thus, a polygraph measures certain physiological variables, for example, pulse rate or blood pressure, in response to stimuli. The measurements may be reliable. Nonetheless, the polygraph is not valid as a lie detector unless the measurements are well correlated with lying.[58]

When there is an established way of measuring a variable, a new measurement process can be validated by comparison with the established one. Breathalyzer readings can be validated against alcohol levels found in blood samples. LSAT or GRE scores used for law school admissions can be validated against grades earned in law school. A common measure of validity is the correlation coefficient between the predictor and the criterion (for example, test scores and later performance).[59]

Employment discrimination cases illustrate some of the difficulties. Plaintiffs suing under Title VII of the Civil Rights Act may challenge an employment

56. David C. Baldus et al., Equal Justice and the Death Penalty: A Legal and Empirical Analysis 49–50 (1990).

57. Alan H. Dorfman & Richard Valliant, *A Re-Analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*, 9 Stat. & Public Pol'y 175 (2022), https://doi.org/10.1080/2330443X.2022.2120137; Hal S. Stern et al., *Reliability and Validity of Forensic Science Evidence*, Significance, Apr. 2019, at 21, 22–23.

58. *See* United States v. Henderson, 409 F.3d 1293, 1303 (11th Cir. 2005) ("while the physical responses recorded by a polygraph machine may be tested, 'there is no available data to prove that those specific responses are attributable to lying'"); Nat'l Research Council, The Polygraph and Lie Detection (2003) (reviewing the scientific literature).

59. As the discussion of the correlation coefficient in the section titled "Correlation Coefficients" below indicates, the closer the coefficient is to 1, the greater the validity. For a review of data on test reliability and validity, see *Measuring Success: Testing, Grades, and the Future of College Admissions* (Jack Buckley et al. eds., 2018).

test that has a disparate impact on a protected group, and defendants may try to justify the use of a test as valid, reliable, and a business necessity.[60] For validation, the most appropriate criterion variable is clear enough: job performance. However, plaintiffs may then challenge the validity of performance ratings. Are they sufficiently related to the actual requirements of the job? Are they biased? As for reliability, plaintiffs would need to show that each measure (the test scores and the later performance ratings) provides consistent measurements.[61]

A further problem is that test-takers are likely to be a select group. The ones who get the jobs are even more highly selected. Generally, selection attenuates (weakens) the correlations because differences in performance within a narrow band of highly qualified applicants do not exhibit as much variation as would be expected if a wider range of applicants were selected.[62] Statistical methods that correct for attenuation depend on assumptions about the nature of the test and the procedures used to select the test-takers; these assumptions may be open to challenge.[63]

Measurements also can be made on a nominal scale, as when a chemist notes that litmus paper has turned red, or a criminalist declares that there is a "physical fit" between two fragments of glass. For binary (yes–no) classifications, validity (accuracy) of the classifier can be evaluated with experiments to estimate "sensitivity" and "specificity" (or corresponding error probabilities). Suppose that firearms examiners are given pairs of spent cartridge cases and are required to decide whether they come from the same gun or from two different guns. The experimenter, who has prepared the test pairs, knows the truth; the examiners do not. Results related to a small experiment are in Table 1.[64]

60. *See, e.g.*, Ricci v. DeStefano, 557 U.S. 557 (2009); Washington v. Davis, 426 U.S. 229, 252 (1976); Albemarle Paper Co. v. Moody, 422 U.S. 405, 430–32 (1975); Griggs v. Duke Power Co., 401 U.S. 424 (1971); Lanning v. S.E. Penn. Transp. Auth., 308 F.3d 286 (3d Cir. 2002).

61. *See* section titled "Is the Measurement Process Reliable?" above (discussing *Hall v. Florida*).

62. For an extreme example, consider a firm that provides personal tutoring for college-entrance examinations and hires only job applicants who had near-perfect scores themselves to be the tutors. It could well be that receiving a high score is associated with being an effective tutor. But if the firm hired no low-scoring applicants as tutors, it would be impossible to see that association in a study of the correlation between the scores of the exclusively high-scoring tutors and those of the students they tutor after being hired.

63. *See* Thad Dunning & David A. Freedman, *Modeling Selection Effects, in* Social Science Methodology 225 (Steven Turner & William Outhwaite eds., 2007); Howard Wainer & David Thissen, *True Score Theory: The Traditional Method, in* Test Scoring 23 (David Thissen & Howard Wainer eds., 2001).

64. The numbers are rounded-off versions of those in Table C1 of Heike Hofmann et al., *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 Law, Probability & Risk 317, 363 (2020), https://doi.org/10.1093/lpr/mgab002 (deducing numbers for independent comparisons within a somewhat differently designed and harder to interpret 2003 experiment with eight FBI examiners). That there are zeros in the cells for false identifications and false eliminations does not mean that these outcomes cannot occur. *See* the "Other situations" subsection for confidence intervals below. Also, in

Table 1. Test Results for Cartridge-case Comparisons

	Same Source	Different Source
Reported Same Source	30	0
Reported Different Source	0	45

The examiners performed flawlessly in these 75 instances. If we call the same-source condition "positive" and the "different-source" condition "negative," there were zero false-positive reports and zero false-negative reports. To put it another way, the examiners were 100% accurate in dealing with same-source pairs—they called 30 out of 30 such pairs positives, for an observed sensitivity of 1; likewise, they were 100% accurate in dealing with different-source pairs—they called 45 out of 45 such pairs different, for an observed specificity of 1. Whether the results of this small experiment are representative of what might be seen in a larger set of experiments, and whether the experiments would be representative of the outcomes in case work are further questions.

Are the Measurements Recorded Correctly?

Judging the adequacy of data collection involves an examination of the process by which measurements are taken. Are responses to interviews coded correctly? Do mistakes distort the results? How much data are missing? What was done to compensate for gaps in the data? These days, data are stored in computer files. Cross-checking the files against the original sources (e.g., paper records), at least on a sample basis, can be informative.

Data quality is a pervasive issue in litigation and in applied statistics more generally.[65] A programmer moves a file from one computer to another, and half the data disappear. The definitions of crucial variables are lost in the sands of time. Values get corrupted: Social Security numbers come to have eight digits instead of nine, and vehicle identification numbers fail the most elementary consistency checks. Everybody in the company, from the CEO to the rawest mailroom trainee,

experiments and in practice, firearms examiners often report comparisons as "inconclusive." This complication is discussed in the section titled "Are the Categories Appropriate?" below.

65. *E.g.*, Bland–Collins v. Howard Univ., 19 F. Supp. 3d 252, 255 (D.D.C. 2014) (statistician who was fired after she discovered "over 5,000 errors in the coded data collected from structured interviews" in a National Science Foundation funded research project brought a whistleblower retaliation claim). Transcription errors put the FBI in the awkward position of using 33 incorrect allele frequencies (out of 1,100) in the software it distributed since 1999 to DNA laboratories for computing random-match probabilities. Tamyra R. Moretti et al., *Erratum*, 60 J. Forensic Sci. 1114 (2015), https://doi.org/10.1111/1556-4029.12806; Spencer S. Hsu, *FBI Notifies Crime Labs of Errors Used in DNA Match Calculations Since 1999*, Wash. Post, May 29, 2015.

turns out to have been hired on the same day. Many of the residential customers have last names that indicate commercial activity (e.g., "Happy Valley Farriers"). These problems seem humdrum by comparison with those of reliability and validity, but—unless caught in time—they can be fatal to statistical arguments.[66]

What Does It Mean to Be Random?

In the law, a selection process sometimes is called "random," provided that it does not exclude identifiable segments of the population. Statisticians use the term in a more rigorous and technical sense. For example, to choose one person at random from a population in the strict statistical sense, we would have to ensure that everybody in the population has the same probability of selection. With a randomized controlled experiment, subjects are assigned to treatment or control at random in the strict sense—by tossing coins, throwing dice, looking at tables of random numbers, or more commonly these days, by using a random number generator on a computer. The same rigorous definition applies to random sampling. Randomness in the technical sense provides assurance of unbiased estimates from a randomized controlled experiment or a probability sample. Randomness in the technical sense also justifies calculations of standard errors, confidence intervals, and *p*-values (see sections titled "What Inferences Can Be Drawn from the Data?" and "Correlation and Regression" below). Looser definitions of randomness are inadequate for statistical purposes.

How Have the Data Been Presented?

After data have been collected, they should be presented in a way that makes them intelligible and that helps reveal their implications. Data can be summarized with a few numbers or with graphical displays. However, the wrong summary can

66. *See, e.g.,* Malletier v. Dooney & Bourke, Inc., 525 F. Supp. 2d 558, 630 (S.D.N.Y. 2007) (coding errors contributed "to the cumulative effect of the methodological errors" that warranted exclusion of a consumer confusion survey); EEOC v. Sears, Roebuck & Co., 628 F. Supp. 1264, 1304, 1305 (N.D. Ill. 1986) (finding that the EEOC "has made so many general coding errors that its data base does not fairly reflect the characteristics of applicants"), *aff'd*, 839 F.2d 302 (7th Cir. 1988); *cf.* EEOC v. Freeman, 961 F. Supp. 2d 783, 796 (D. Md. 2013) (excluding an EEOC analysis of an employer's records purporting to show disparate impact of credit and criminal records checks of job applicants because the EEOC database was not a random sample but rather was "cherry-picked" and "[t]he mind-boggling number of errors contained in [its] database could alone render [the] conclusions worthless").

mislead.[67] The section "Are Rates or Percentages Properly Interpreted?" below discusses rates or percentages and provides some cautionary examples of misleading summaries, indicating the kinds of questions that might be considered when summaries are presented in court. Percentages are often used to demonstrate statistical association, which is the topic of the section titled "Is an Appropriate Measure of Association Used?" The section titled "Does a Graph Portray Data Fairly?" considers graphical summaries of data, while the sections titled "Is an Appropriate Measure Used for the Center of a Distribution?" and "Is an Appropriate Measure of Variability Used?" discuss some of the basic descriptive statistics that are likely to be encountered in litigation, including the mean, median, and standard deviation.

Are Rates or Percentages Properly Interpreted?

How Big Is the Base of a Percentage?

Rates and percentages often provide effective summaries of data, but these statistics can be misinterpreted. A rate reports a comparison of one number against some other quantity, for example, the number of reported crimes per hundred thousand residents. A percentage reports a comparison between two numbers by putting them in terms of a common base (100). Expressing the ratio of the two numbers on a common base makes it easy to compare them.

One application of percentages is for reporting increases or decreases in a quantity or rate by describing the percent change compared to the initial or base amount. When the base is small, however, a small change in absolute terms can generate a large percentage gain or loss. (This could lead to newspaper headlines such as "Increase in Rate of Thefts Alarming," even when the total number of thefts is small.[68]) Conversely, a large base will make for small percentage increases. In these situations, actual numbers may be more revealing than percentages.

Have Appropriate Benchmarks Been Provided?

The selective presentation of numerical information is like quoting someone out of context. Is the fact that a particular actively managed fund of large-cap stocks boasted a return of 25% in 2021 indicative of outstanding management? Considering that the 500 large-cap stocks in "the benchmark S&P 500 notched a total

67. *See generally* Freedman et al., *supra* note 14; Huff, *supra* note 14; Moore & Notz, *supra* note 14; Zeisel, *supra* note 14.

68. Lyda Longa, *Increase in Thefts Alarming*, Daytona News-J. June 8, 2008 (reporting a 35% increase in armed robberies in Daytona Beach, Florida, in a 5-month period, but not indicating whether the number had gone up by 6 (from 17 to 23), by 300 (from 850 to 1,150), or by some other amount).

return . . . of 28.7% [that] year," a growth rate of 25% is less indicative of unusual financial acumen than one might have thought.[69] In this example and many others, it is helpful to find a benchmark that puts the figures into perspective.[70]

Have the Data Collection Procedures Changed?

Changes in the process of collecting data can create problems of interpretation. Statistics on crime provide many examples. The number of petty larcenies reported in Chicago more than doubled one year—not because of an abrupt crime wave, but because a new police commissioner introduced an improved reporting system.[71] For a time, police officials in Washington, D.C., "demonstrated" the success of a law-and-order campaign by valuing stolen goods at $49, just below the $50 threshold then used for inclusion in the FBI's Uniform Crime Reports.[72] Allegations of manipulation in the reporting of crime from one time period to another are legion.[73] Almost all series of numbers that cover many years are affected by changes in definitions and collection methods. When a study includes such time-series data, it is useful to inquire about changes in the way data are collected or reported and to look for any sudden jumps, which may signal such changes.

69. Karen Langley, *Stock Pickers Watched the S&P 500 Pass Them By Again in 2021*, Wall St. J., Mar. 16, 2022 (reporting that "[t]he failure of stock pickers to beat the benchmark is nothing new: 2021 was the 12th consecutive year in which the majority of actively-managed funds of large-cap stocks watched the S&P 500 pass them by").

70. The selection of the benchmark may merit scrutiny. Securities and Exchange Commission (SEC) Rule 33–698 requires mutual funds to display past returns alongside those of "an appropriate broad-based securities market index." But the rule "does not prohibit funds from comparing their past returns to those of newly-chosen index(es)." Kevin Mullally & Andrea Rossi, *Moving the Goalposts? Mutual Fund Benchmark Changes and Performance Manipulation*, June 24, 2022, https://perma.cc/6MFU-WYEN. There is evidence that, to attract more investors, funds simply change their benchmark indexes to lower performing ones—including benchmarks that are not the best match for the funds' actual investment strategies. *Id.*

71. James P. Levine et al., Criminal Justice in America: Law in Action 99 (1986) (referring to a change from 1959 to 1960).

72. David Seidman & Michael Couzens, *Getting the Crime Rate Down: Political Pressure and Crime Reporting*, 8 Law & Soc'y Rev. 457 (1974).

73. John A. Eterno & Eli B. Silverman, The Crime Numbers Game: Management by Manipulation (2012); Michael D. Maltz, *Missing UCR Data and Divergence of the NCVS and UCR Trends*, *in* Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR 269, 280 (James P. Lynch & Lynn A. Addington eds., 2006), https://doi.org/10.1017/CBO9780511618543.010 (citing newspaper reports in Boca Raton, Atlanta, New York, Philadelphia, Broward County (Florida), and St. Louis). Changes in reporting practices also can be improvements, but they can still distort comparisons to data collected before the changes. *E.g.*, Greg Moreau, Statistics Canada, Police-Reported Crime Statistics in Canada 2018, at 7 (2019), https://perma.cc/AR6J-DFTF.

Are the Categories Appropriate?

Misleading summaries also can be produced by the categories used for comparison. In *Philip Morris, Inc. v. Loew's Theatres, Inc.*,[74] and *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*,[75] Philip Morris and R.J. Reynolds sought an injunction to stop the maker of Triumph low-tar cigarettes from running advertisements claiming that participants in a national taste test preferred Triumph to other brands. Plaintiffs alleged that claims that Triumph was a "national taste test winner" or Triumph "beats" other brands were false and misleading. An exhibit introduced by the defendant contained the data shown in Table 2.[76] Only 14% + 22% = 36% of the sample preferred Triumph to Merit, whereas 29% + 11% = 40% preferred Merit to Triumph. By selectively combining categories, however, the defendant attempted to create a different impression. Because 24% found the brands to be about the same, and 36% preferred Triumph, the defendant claimed that a clear majority (36% + 24% = 60%) found Triumph "as good [as] or better than Merit."[77] The court resisted this chicanery, finding that defendant's test results did not support the advertising claims.[78]

Table 2. Data Used by a Defendant to Refute Plaintiff's False Advertising Claims

	Triumph Much Better	Triumph Somewhat Better	Triumph About the Same	Triumph Somewhat Worse	Triumph Much Worse
Number	45	73	77	93	36
Percentage	14	22	24	29	11

There was a similar distortion in claims for the accuracy of a home pregnancy test. The manufacturer advertised the test as 99.5% accurate under laboratory conditions. The underlying data are summarized in Table 3.

Table 3. Home Pregnancy Test Results

	Actually Pregnant	Actually Not Pregnant
Test says pregnant	197	0
Test says not pregnant	1	2
Total	198	2

74. 511 F. Supp. 855 (S.D.N.Y. 1980).
75. 511 F. Supp. 867 (S.D.N.Y. 1980).
76. *Philip Morris*, 511 F. Supp. at 866.
77. *Id.*
78. *Id.* at 856–57.

The table does indicate that only one error occurred in 200 assessments, for 99.5% overall accuracy. But the table also shows that the test can make two types of errors: It can tell a pregnant woman that she is not pregnant (a false negative), and it can tell a woman who is not pregnant that she is (a false positive). The reported 99.5% accuracy rate conceals a crucial fact—the company had virtually no data with which to measure the rate of false positives.[79]

The problem with combining categories into broader ones has surfaced in criminal cases as well. As noted in the section titled "Is the Measurement Process Valid?" above, criminalists examine pairs of items, such as spent cartridge cases, to decide whether they are associated with the same source. In firearms-toolmark matching, the traditional comparison between an item of known origin and the item whose origin is in question culminates in a report of "identification," "inconclusive," or "elimination." In validity studies, toolmark examiners are given items to evaluate when the experimenter, but not the criminalist, knows whether the items are from the same source or from two different sources. Table 4 expands Table 1 with a row for "inconclusives" in the experiment:

Table 4. Test Results for Cartridge-case comparisons[80]

	Same Source	Different Source
Reported Same Source	30	0
Reported Different Source	0	45
Reported Inconclusive	0	157

The new row reveals that the criminalists reached no definitive conclusion in most of the cases, and that all these "inconclusives" pertained to pairs of cartridge cases from two different guns. Adding in all these "inconclusives," the overall proportion of correct decisions drops from 100% to only $(30 + 45) / (30 + 45 + 157) = 32\%$.

79. Only two women in the sample were not pregnant; the test gave correct results for both of them (a specificity of 100%). Although a false-positive rate of 0 is ideal, an estimate based on a sample of only two women is not. These data are reported in Arnold Barnett, *How Numbers Can Trick You*, Tech. Rev., Oct. 1994, at 38, 44–45. When data on test performance are sufficient to accurately estimate (1) the probability of a positive when the condition is present (sensitivity) and (2) the probability of a negative when it is not (specificity), the estimates can be combined to express the probative value of a test result in the form of a likelihood ratio. *See* Appendix section titled "What Is Bayes' Rule?" below. Ideally, the test should have both high sensitivity and specificity. However, the combination of moderate sensitivity and high specificity also indicates that a positive finding strongly supports the conclusion that the condition is present. *Id.*

80. The numbers are similar to those in Hofmann et al., *supra* note 64, at 363 (tbl. C1). More recent and larger studies with different results also are noted there.

In one sense, this is a poor "correct decision rate."[81] It indicates that examiners are missing many opportunities to reach conclusions; perhaps, if they were more venturesome, they would produce more evidence correctly excluding or correctly implicating suspects. Or perhaps they would simply make more mistakes. In any event, as with the advertised accuracy rate of 99.5% for the home pregnancy test, the overall correct-decision rate masks a crucial fact—here, that the examiners in the study, when confronted with marks from different guns, never erred in attributing them to the same gun. The low false-positive rate (0/45 incorrect same-source decisions) is more pertinent to a same-source determination offered to implicate the defendant than is the aggregated correct-decision rate.[82] However, observing relatively few false positives (or even none at all, as in this study) may not be convincing—particularly with small or unrepresentative samples—and better measures for the probative value of a positive test result are available.[83]

What Comparisons Are Made?

Finally, there is the issue of which numbers to compare. Researchers sometimes choose among alternative comparisons. Why did they choose the one they did?

81. *See* People v. Winfield, No. 15 CR 14066–01 (Ill. Cir. Ct. Feb. 9, 2023) (aff. of Susan Vanderplas, Kori Khan, Heike Hofmann, Alicia Carriquiry, Jan. 3, 2022) (affidavit submitted in murder case characterizing the "correct source decision rates" in experiments comparable to the one described here as "abysmal").

82. Some scientists and litigants maintain that, by definition, every inconclusive result is an error (and should be counted as such) because it does not express the true association (positive or negative) within each pair. *Id.* The opposing view is that "inconclusive" is more of a conservative decision to opt-out of the binary classification scheme shown in Table 1 than an inference to the true state of affairs. Under this view, forensic scientists and policy makers might wish to investigate whether examiners should be more willing to make inclusions or exclusions (or to describe the strength of the evidence on a more graduated scale). But a report of "I cannot tell" is neither an erroneous nor a correct statement about the true source. This understanding has led commentators to argue that the number of "inconclusives" does not belong in either the numerator or the denominator of the proportions used to assess the validity of the positive or negative findings of association. *See, e.g.,* Hal R. Arkes & Jonathan J. Koehler, *Inconclusives and Error Rates in Forensic Science: A Signal Detection Theory Approach*, 20 Law, Probability & Risk 153 (2021), https://doi.org/10.1093/lpr/mgac005. In response, it has been said that most inconclusives in validity studies would be false inclusions in practice, which can render an error rate that does not incorporate inconclusives misleading. In short, a range of views currently exists regarding what to count as "errors" in validity studies when extrapolating to actual casework. *See* Hal R. Arkes & Jonathan J. Koehler, *Inconclusive Conclusions in Forensic Science: Rejoinders to Scurich, Morrison, Sinha and Gutierrez*, 21 Law, Probability & Risk 175 (2022), https://doi.org/10.1093/lpr/mgad002.

83. We discuss sampling error in false-positive proportions in the section titled "What Is the Standard Error?" below and describe a more complete measure of probative value in the Appendix section titled "What Is Bayes' Rule?"

Would another comparison give a different view? A government agency, for example, may want to compare the amount of service now being given with that of earlier years—but what earlier year should be the baseline? If the first year of operation is used, a large percentage increase should be expected because of startup problems. If last year is used as the base, was it also part of the trend, or was it an unusually poor year? If the base year is not representative of other years, the percentage may not portray the trend fairly. No single question can be formulated to detect such distortions, but it may help to ask for the numbers from which the percentages were obtained; asking about the base can also be helpful.[84]

Is an Appropriate Measure of Association Used?

Many cases involve statistical association. Does a test for employee promotion have an exclusionary effect that depends on race or sex? Does the incidence of murder vary with the rate of executions for convicted murderers? Do consumer purchases of a product depend on the presence or absence of a product warning? This section discusses tables and percentage-based statistics that are frequently presented to answer such questions.[85]

Percentages often are used to describe the association between two variables. Suppose that a university is alleged to discriminate against women in admitting students, and that the university consists of only two colleges—engineering and business. The university admits 350 out of 800 male applicants; by comparison, it admits only 200 out of 600 female applicants. Such data commonly are displayed as in Table 5.[86]

Table 5. Admissions by Sex

Decision	Male	Female	Total
Admit	350	200	550
Deny	450	400	850
Total	800	600	1,400

The table indicates that 350/800 = 44% of the males are admitted, compared with only 200/600 = 33% of the females. One way to express the disparity is to subtract the two percentages: 44% – 33% = 11 percentage points. Although such

84. For assistance in coping with percentages, see Zeisel, *supra* note 14, at 1–24.

85. Correlation and regression are discussed in the section titled "Correlation and Regression" below.

86. A table of this sort is called a "cross-tab" or a "contingency table." Table 5 is "two-by-two" because it has two rows and two columns, not counting rows or columns containing totals.

subtraction is commonly seen in jury discrimination cases,[87] the difference is inevitably small when the two percentages are both close to zero. If the selection rate for males is 5% and that for females is 1%, the difference is only 4 percentage points. Yet, females have only one-fifth the chance of males of being admitted, and that may be of real concern.

For Table 5, the selection ratio (used by the Equal Employment Opportunity Commission in its "80% rule") is 33/44 = 75%, meaning that, on average, women have 75% the chance of admission that men have.[88] The analogous statistic used in epidemiology is called the relative risk.[89] Relative risks are usually quoted as decimals; for example, a selection ratio of 75% corresponds to a relative risk of 0.75.

However, the selection ratio has its own problems. In the last example, if the selection rates are 5% and 1%, then the exclusion rates are 95% and 99%. The ratio is 99/95 = 104%, meaning that females have, on average, 104% the risk of males of being rejected. The underlying facts are the same, of course, but this formulation sounds much less disturbing.

A statistic known as the odds ratio is more symmetric. If 5% of male applicants are admitted, the odds on a man being admitted are 5 to 95 = 1 to 19; the odds on a woman being admitted are 1:99. The odds ratio is (1:99)/(1:19) = 19:99. The odds ratio for rejection instead of acceptance is the same, except that the order is reversed.[90] Although the odds ratio has desirable mathematical properties, its meaning may be less clear than that of the selection ratio or the simple difference.[91]

Data showing disparate impact are generally obtained by aggregating—putting together—data from a variety of sources. Unless the source material is fairly homogeneous, aggregation can distort patterns in the data. We illustrate

87. *E.g.*, Woodfox v. Cain, 772 F.3d 358, 376 (5th Cir. 2014) (presenting percentage-point differences that were seen as establishing a prima facie case of discrimination); Sara Sun Beale, Grand Jury Law and Practice §§ 3:18–3:19 (2d ed. update Dec. 2021); David H. Kaye, *Statistical Evidence of Discrimination in Jury Selection*, *in* Statistical Methods in Discrimination Litigation 13 (David H. Kaye & Mikel Aickin eds., 1986).

88. A procedure that selects candidates from the least successful group at a rate less than 80% of the rate for the most successful group "will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (1978). The rule is designed to help spot instances of substantially discriminatory practices, and the commission usually asks employers to justify any procedures that produce selection ratios of 80% or less.

89. *See* Gold et al., *supra* note 15.

90. For women, the odds of rejection are 99 to 1; for men, 19 to 1. The ratio of these odds is 99:19. Likewise, the odds ratio for an admitted applicant being a man as opposed to a denied applicant being a man is 99:19.

91. *But see* Joseph B. Kadane, *Odds Ratios as a Measure of Disproportionate Treatment: Application to Jury Venires*, 21 Law, Probability & Risk 163, 172 (2022), https://doi.org/10.1093/lpr/mgad003 ("[o]dds ratios are a natural and easily interpretable way of summarizing data in cases alleging racial disproportion").

the problem with the hypothetical admission data in Table 5. Applicants can be classified not only by gender and admission but also by the college to which they applied, as in Table 6.

Table 6. Admissions by Sex and College

	Engineering		Business	
Decision	Male	Female	Male	Female
Admit	300	100	50	100
Deny	300	100	150	300

The entries in Table 6 add up to the entries in Table 5. Expressed in a more technical manner, Table 5 is obtained by aggregating the data in Table 6. Yet there is no association between sex and admission in either college; within each college, men and women are admitted at identical rates. Combining two colleges with no association produces a university in which sex is associated strongly with admission. The explanation for this paradox is that the business college, to which most of the women applied, admits relatively few applicants. It is easier to be accepted at the engineering college, the college to which most of the men applied. Combining data from heterogeneous sources (two very different colleges) has made the selectivity of the college into a confounding variable.[92]

Does a Graph Portray Data Fairly?

Graphs are useful for revealing key characteristics of a batch of numbers, trends over time, and the relationships among variables.[93]

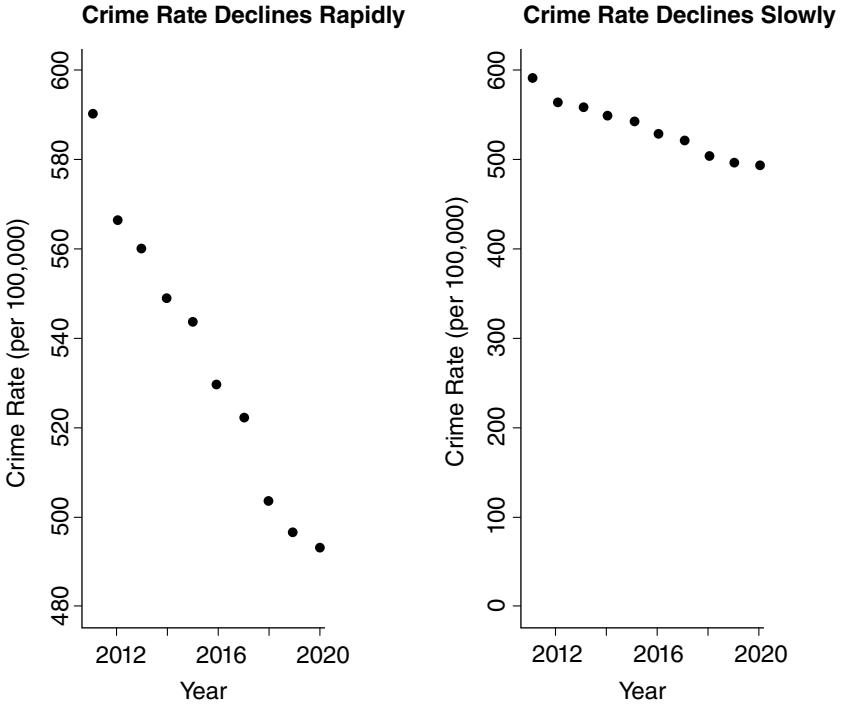
How Are Trends Displayed?

Graphs that plot values over time are useful for seeing trends. However, the scales on the axes matter. In Figures 1 and 2, the rate of all crimes of domestic violence in Florida (per 100,000 people) appears to decline rapidly over the 10 years from

92. Tables 5 and 6 are hypothetical, but closely patterned on a real example. *See* P.J. Bickel et al., *Sex Bias in Graduate Admissions: Data from Berkeley*, 187 Science 398 (1975), https://doi.org/10.1126/science.197.4175.398. The tables are an instance of Simpson's paradox.

93. *See generally* Alberto Cairo, The Truthful Art: Data, Charts, and Maps for Communication (2016); Alberto Cairo, The Functional Art: An Introduction to Information Graphics and Visualization (2012); William S. Cleveland, The Elements of Graphing Data (rev. ed. 1994); Edward R. Tufte, The Visual Display of Quantitative Information (2d ed. 2001).

Figures 1 and 2. Manipulating the scale of a graph.



2011 through 2020; in Figure 2, the same rate appears to drop slowly.[94] The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate. A similar admonition applies to bar charts, which display counts or rates across categories.[95]

How Are Distributions Displayed?

A graph commonly used to display the distribution of data is the histogram. One axis denotes the numbers, and the other indicates how often these numbers fall within specified intervals (called "bins" or "class intervals"). For example, we

94. Florida Statistical Analysis Center, Florida Dep't of Law Enforcement, Statewide Reported Domestic Violence Offenses in Florida, 1992–2020, https://perma.cc/KQ7F-H6UM. The data are from the Florida Uniform Crime Report statistics on crimes ranging from simple stalking and forcible fondling to murder and arson.

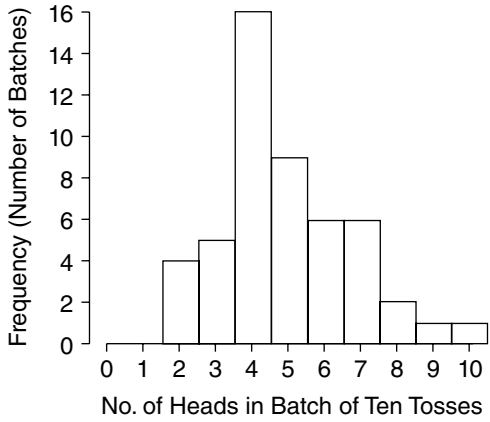
95. *See* David Spiegelhalter, The Art of Statistics: Learning from Data 25–26 (2019) (for an example using survival rates of patients admitted to different hospitals).

flipped a quarter 10 times in a row and counted the number of heads in this "batch" of 10 tosses. Repeating this exercise until we obtained 50 batches, we recorded the following counts:[96]

7 7 5 6 8	4 2 3 6 5	4 3 4 7 4	6 8 4 7 4	7 4 5 4 3
4 4 2 5 3	5 4 2 4 4	5 7 2 3 5	4 6 4 9 10	5 5 6 6 4

The histogram is shown in Figure 3.[97] A histogram displays how the data are distributed over the range of possible values. The spread can be made to appear larger or smaller, however, by changing the scale of the horizontal axis. Likewise, the shape can be altered somewhat by changing the size of the bins. It may be worth inquiring how the analyst chose the bin widths. As the width of the bins decreases, the graph becomes more detailed, but the appearance becomes

Figure 3. Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter.



96. The coin landed heads 7 times in the first 10 tosses; by coincidence, there were also 7 heads in the next 10 tosses; there were 5 heads in the third batch of 10 tosses; and so forth.

97. In Figure 3, the bin width is 1. There were no 0's or 1's in the data, so the bars over 0 and 1 disappear. There is a bin from 1.5 to 2.5; the four 2's in the data fall into this bin, so the bar over the interval from 1.5 to 2.5 has height 4. There is another bin from 2.5 to 3.5, which catches five 3's; the height of the corresponding bar is 5. And so forth. Five is the most likely count in any random batch, but in these 50 batches, a count of 4 heads occurred more frequently, as shown by the larger height for bin 4.

All the bins in Figure 3 have the same width, so this histogram is just like a bar graph. However, data are often published in tables with unequal intervals. The resulting histograms will have unequal bin widths; bar heights should be calculated so that the areas (height × width) are proportional to the frequencies. In general, a histogram differs from a bar graph in that it represents frequencies by area, not height. *See* Freedman et al., *supra* note 14, at 31–41.

more ragged until finally the graph is effectively a plot of each datum. The optimal bin width depends on the subject matter and the goal of the analysis.

Is an Appropriate Measure Used for the Center of a Distribution?

Perhaps the most familiar descriptive statistic is the mean (or "arithmetic mean"). The mean can be found by adding all the numbers and dividing the total by how many numbers were added. By comparison, the median cuts the numbers into halves: half the numbers are larger than the median and half are smaller.[98] Yet a third statistic is the mode, which is the most common number in the dataset. These statistics are different, although they are not always clearly distinguished. In ordinary language, the arithmetic mean, the median, and the mode seem to be referred to interchangeably as "the average." In statistical parlance, however, the average is the arithmetic mean. The mode is rarely used by statisticians, because it is unstable: Small changes to the data often result in large changes to the mode. The mean takes account of all the data—it involves the total of all the numbers; however, particularly with small datasets, a few unusually large or small observations may have too much influence on the mean. The median is resistant to such outliers.

Studies of damage awards in tort cases find that the mean is larger than the median.[99] This is because the mean takes into account (indeed, is heavily influenced by) the magnitudes of the relatively few very large awards, whereas the median is influenced only by the number of such awards. If one is seeking a single, representative number for the awards, the median may be more useful than the mean.[100] Still, if the issue is whether insurers were experiencing more

98. Technically, at least half the numbers are at the median or larger; at least half are at the median or smaller. When the distribution is symmetric, the mean equals the median. The values diverge, however, when the distribution is asymmetric, or skewed.

99. Herbert M. Kritzer et al., *An Exploration of "Noneconomic" Damages in Civil Jury Awards*, 55 Wm. & Mary L. Rev. 971 (2014) (summarizing empirical research); Thomas H. Cohen & Steven K. Smith, U.S. Dep't of Justice, Bureau of Justice Statistics Bulletin NCJ 202803, Civil Trial Cases and Verdicts in Large Counties 2001, 10 (2004) (in a probability sample of cases, the median compensatory award in wrongful death cases was $961,000, whereas the mean award was around $3.75 million for the 162 cases in which the plaintiff prevailed); *cf.* Stephen J. Choi & Theodore Eisenberg, *Punitive Damages in Securities Arbitration: An Empirical Study*, 39 J. Legal Stud. 497, 513 tbl. 1 (2010) (reporting much higher mean than median arbitration awards). In *TXO Production Corp. v. Alliance Resources Corp.*, 509 U.S. 443 (1993), briefs portraying the punitive damage system as out of control pointed to mean punitive awards. These were some 10 times larger than the median awards described in briefs defending the system of punitive damages. Michael Rustad & Thomas Koenig, *The Supreme Court and Junk Social Science: Selective Distortion in Amicus Briefs*, 72 N.C. L. Rev. 91, 145–47 (1993).

100. *E.g.*, Heinrich v. Sweet, 308 F.3d 48 (1st Cir. 2002) (two outliers made the median survival time a better indicator of what might have happened had a patient not undergone

costs from jury verdicts, the mean is the more appropriate statistic: The total of the awards is directly related to the mean, not to the median.[101]

Research also has shown considerable stability in the ratio of punitive to compensatory damage awards, and the Supreme Court has placed great weight on this ratio in deciding whether punitive damages are excessive in a particular case. In *Exxon Shipping Co. v. Baker*,[102] Exxon contended that an award of $2.5 billion in punitive damages for a catastrophic oil spill in Alaska was unreasonable under federal maritime law. The Court looked to a "comprehensive study of punitive damages awarded by juries in state civil trials [that] found a median ratio of punitive to compensatory awards of just 0.62:1, but a mean ratio of 2.90:1."[103] The higher mean could be the result of some large and atypical punitive awards.[104] Looking to the median ratio as "the line near which cases like this one largely should be grouped," the majority concluded that "a 1:1 ratio, which is above the median award, is a fair upper limit in such maritime cases [of reckless conduct]."[105]

Is an Appropriate Measure of Variability Used?

The location of the center of a batch of numbers reveals nothing about the variations exhibited by these numbers. The numbers 1, 2, 5, 8, 9 have 5 as their mean and median. So do the numbers 5, 5, 5, 5, 5. In the first batch, the numbers vary considerably about their mean; in the second, the numbers do not vary at all.

experimental therapy). In passing on proposed settlements in class-action lawsuits, courts have been advised to look to the magnitude of the settlements negotiated by the parties. But the mean settlement will be large if a higher number of meritorious, high-cost cases are resolved early in the life cycle of the litigation. This possibility led the court in *In re Educational Testing Service Praxis Principles of Learning and Teaching, Grades 7–12 Litigation*, 447 F. Supp. 2d 612, 625 (E.D. La. 2006), to regard the smaller median settlement as "more representative of the value of a typical claim than the mean value" and to use this median in extrapolating to the entire class of pending claims.

101. To get the total award, just multiply the mean by the number of awards; by contrast, the total cannot be computed from the median. (The more pertinent figure for the insurance industry is not the total of jury awards, but actual claims experience including settlements; of course, even the risk of large punitive damage awards may have considerable impact.)

102. 554 U.S. 471 (2008).

103. *Id.* at 499.

104. According to the Court, "the outlier cases subject defendants to [disproportionate] punitive damages," *id.* at 500, and the "stark unpredictability" of these rare awards is the "real problem." *Id.* at 499. This perceived unpredictability has been the subject of various statistical studies and much debate. *See* Theodore Eisenberg et al., *Variability in Punitive Damages: Empirically Assessing* Exxon Shipping Co. v. Baker, 166 J. Institutional & Theoretical Econ. 5 (2010) (also criticizing the use of a constant 1:1 ratio across the entire range of damage awards); Anthony J. Sebok, *Punitive Damages: From Myth to Theory*, 92 Iowa L. Rev. 957 (2007).

105. 554 U.S. at 513.

Statistical measures of variability include the range, the interquartile range, and the standard deviation. The range is the difference between the largest number in the batch and the smallest. The range seems natural, and it indicates the maximum spread in the numbers, but the range is unstable because it depends entirely on the most extreme values.

The interquartile range is the difference between the 25th and 75th percentiles. By definition, 25% of the data fall below the 25th percentile, 90% fall below the 90th percentile, and so on. The median is the 50th percentile. The interquartile range contains 50% of the numbers and is resistant to changes in extreme values.

The standard deviation can be viewed as a kind of average or typical deviation from the mean.[106] Suppose we have a batch of 50 numbers whose mean is 100. The standard deviation is found by subtracting this mean from each number, squaring each of these deviations, adding up the squared deviations, dividing by 50 (to get the mean squared deviation, or "variance"), and taking the square root. For example, if one of the numbers is 105, then it deviates from the mean by 5, and the square of 5 is $5^2 = 25$. If the mean of all such squared deviations is 400, then the standard deviation is the square root of 400, which is 20. Taking the square root gets back to the original scale of the measurements. For example, if the numbers are measurements of length in inches, the mean and standard deviation are also in inches.

There are no hard and fast rules about which statistic is the best. In general, the bigger the measures of spread are, the more the numbers are dispersed.[107] Particularly in small datasets, the standard deviation can be influenced heavily by a few outlying values. To assess the extent of this influence, the mean and the standard deviation can be recomputed with the outliers discarded. These "trimmed" statistics (and some others) are more robust (less sensitive to outliers). Beyond this, any of the statistics can (and often should) be supplemented with a diagram that displays much of the data.

106. When the distribution follows the normal curve, about 68% of the data will lie within plus-or-minus 1 standard deviation of the mean; about 95% will lie within 2 standard deviations of the mean. For other distributions, the proportions will be different.

107. In *Exxon Shipping Co. v. Baker*, 554 U.S. 471 (2008), along with the mean and median ratios of punitive to compensatory awards of 0.62 and 2.90, the Court referred to a standard deviation of 13.81. *Id.* at 499. These numbers led the Court to remark that "[e]ven to those of us unsophisticated in statistics, the thrust of these figures is clear: the spread is great, and the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories." *Id.* at 500. The size of the standard deviation compared to the mean supports the observation that ratios in the cases of jury award studies are dispersed. A graph of each pair of punitive and compensatory damages offers more insight into how scattered these figures are. *See* Theodore Eisenberg et al., *The Predictability of Punitive Damages*, 26 J. Legal Stud. 623 (1997), and the section titled "Scatter Diagrams" below.

What Inferences Can Be Drawn from the Data?

The inferences that may be drawn from a study depend on the design of the study and the quality of the data (see section titled "How Have the Data Been Collected?" above). The data might not address the issue of interest, might be systematically in error, or might be difficult to interpret because of confounding. Statisticians would group these concerns together under the rubric of "bias." In this context, bias means systematic error, with no connotation of prejudice. We turn now to another concern, namely, the impact of random chance on study results. This contribution to total error may be called random error, sampling error, chance error, or statistical error.[108]

If a pattern in the data is the result of chance, it is likely to wash out when more data are collected. By applying the laws of probability, a statistician can assess the likelihood that random error will create spurious patterns of certain kinds. Such assessments are often viewed as essential when making inferences from data. Thus, statistical inference typically involves tasks such as the following, which will be discussed in the rest of this guide.

- *Estimation.* A statistician draws a sample from a population (see section titled "Descriptive Surveys and Censuses" above) and estimates a parameter—that is, a numerical characteristic of the population. Random error will throw the estimate off the mark. The question is, by how much? The precision of an estimate is usually reported in terms of the standard error and a confidence interval.
- *Significance testing.* A "null hypothesis" is formulated—for example, that a parameter takes a particular value. Because of random error, an estimated value for the parameter is likely to differ from the value specified by the null—even if the null is right. ("Null hypothesis" is often shortened to "null.") How likely is it to get a difference as large as, or larger than, the one observed in the data? This chance is known as a *p*-value. Small *p*-values argue against the null hypothesis as such values suggest the observed difference is not likely due to chance alone. Statistical significance can be determined by reference to the *p*-value; significance testing is the technique for computing *p*-values and determining statistical significance. Significance testing is sometimes called hypothesis testing; some statisticians

108. Econometricians use the parallel concept of random disturbance terms. *See* Rubinfeld & Card, *supra* note 25. Randomness and cognate terms have precise technical meanings; it is randomness in the technical sense that justifies the probability calculations behind standard errors, confidence intervals, and *p*-values (see section titled "What Does It Mean to Be Random?" above and the sections titled "Estimation" and "*p*-values, Significance Levels, and Hypothesis Tests" below). For a discussion of samples and populations, see section titled "Descriptive Surveys and Censuses" above.

restrict the latter term to instances in which there are two explicit hypotheses to be addressed.[109]

- *Developing a statistical model.* Statistical inferences often depend on the validity of statistical models for the data. If the data are collected on the basis of a probability sample or a randomized experiment, there will be statistical models that suit the occasion, and inferences based on these models will be secure. Otherwise, calculations are generally based on analogy: This group of people is like a random sample; that observational study is like a randomized experiment. The fit between the statistical model and the data-collection process may then require examination—how good is the analogy? If the model breaks down, that will bias the analysis.
- *Computing posterior probabilities.* Given the sample data, what is the probability that a particular hypothesis about the population (such as the null hypothesis in a significance test) is true? The question might be of direct interest to the courts, especially when translated into English; for example, the null hypothesis might be that African-American and white job applicants have the same probability of passing a qualifying examination—in other words, the test does not adversely impact one group. Posterior probabilities can be computed using a formula called Bayes' rule.[110]

109. As explained in the section titled "*p*-values, Significance Levels, and Hypothesis Tests" below, tests of significance focus on the degree to which data are inconsistent with a specified null hypothesis. There is no explicit alternative hypothesis, although frequently an alternative is implicit in the choice to use a one-sided or two-sided test. The *p*-value is the probability that data as or more unexpected than the data observed would occur by chance under the null hypothesis. The statistical significance of the result is assessed in terms of the *p*-value. Sir Ronald Fisher usually is credited with introducing this framework. *But see* Michael Cowles & Caroline Davis, *On the Origins of the .05 Level of Statistical Significance*, 37 Am. Psych. 553 (1982).

By contrast, formal hypothesis testing (as developed by Jerzy Neyman and Egon Pearson) requires explicit specification of an alternative hypothesis and then development of a test procedure with pre-specified probabilities of making two types of errors—type I (rejecting the null hypothesis when it is true) and type II (failing to reject the null hypothesis when the alternative is true). The end result here is a decision rather than a *p*-value. The two procedures end up being closely related in practice despite their philosophical differences. *See, e.g.,* Erich L. Lehmann, *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 J. Am. Stat. Ass'n 1242 (1993), https//doi.org/10.2307/2291263.

110. The theorem is named after the Reverend Thomas Bayes (England, c. 1701–1761). An elementary version of the rule is derived in the Appendix. Bayes' essay on the subject was published after his death: *An Essay Toward Solving a Problem in the Doctrine of Chances*, 53 Phil. Trans. Royal Soc'y London 370 (1763–1764). On the foundations and varieties of Bayesian and other forms of statistical inference, see, for example, Richard M. Royall, *Statistical Inference: A Likelihood Paradigm* (1997); David Freedman, *Some Issues in the Foundation of Statistics*, 1 Found. Sci. 19 (1995), *reprinted in* Topics in the Foundation of Statistics 19 (Bas C. van Fraasen ed., 1997); Hal S. Stern, *Comparing Philosophies of Statistical Inference*, *in* Handbook of Forensic Statistics 91 (David Banks et al. eds., 2021); see also David H. Kaye, *What Is Bayesianism? in* Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism (Peter Tillers & Eric Green eds., 1988), *reprinted in*

Key ideas of estimation and testing will be illustrated by courtroom examples, with some complications and mathematical details omitted for ease of presentation.[111] Bayesian reasoning with regard to forensic-science testimony is discussed in the Appendix.

The first example, on estimation, concerns the Presidential Recordings and Materials Preservation Act of 1974, which impounded President Richard Nixon's presidential papers after he resigned.[112] Nixon sued, seeking compensation on the theory that the materials belonged to him personally. Courts ruled in his favor: Nixon was entitled to the fair market value of the papers, with the amount to be proved at trial.[113]

The Nixon papers were stored in 20,000 boxes at the National Archives in Alexandria, Virginia. It was plainly impossible to value this entire population of material. Appraisers for the plaintiff therefore took a random sample of 500 boxes. (From this point on, details are simplified; thus, the example becomes somewhat hypothetical.) The appraisers determined the fair market value of each sample box. The average of the 500 sample values turned out to be $2,000. The standard deviation (see section titled "Is an Appropriate Measure of Variability Used?" above) of the 500 sample values was $2,200. Many boxes had low appraised values, whereas some boxes were considered to be extremely valuable; this spread explains the large standard deviation.

Estimation

What Estimator Should Be Used?

With the Nixon papers, it is natural to use the average value of the 500 sample boxes to estimate the average value of all 20,000 boxes comprising the population. With the average value for each box having been estimated as $2,000, the plaintiff demanded compensation in the amount of $20{,}000 \times \$2{,}000 = \$40{,}000{,}000$.

In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller (or less costly) errors are preferred; however, "error" (or the losses that result from errors) might be quantified in more than one way. Moreover, the advantage of one estimator over another may depend on features of the population that are largely unknown, at least before the data are collected and analyzed. For complicated problems,

28 Jurimetrics J. 161 (1988) (distinguishing between "Bayesian probability," "Bayesian statistical inference," "Bayesian inference writ large," and "Bayesian decision theory").

111. Some technical details appear in the appendices to David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* Reference Manual on Scientific Evidence (3d ed. 2011).

112. 44 U.S.C. § 2111.

113. Nixon v. United States, 978 F.2d 1269 (D.C. Cir. 1992); Griffin v. United States, 935 F. Supp. 1 (D.D.C. 1995).

professional skill and judgment may therefore be required when choosing a sample design and an estimator. In such cases, the choices and the rationale for them should be documented.

What Is the Standard Error?

An estimate based on a sample is likely to be off the mark, at least by a small amount, because of random error. The standard error gives the likely magnitude of this random error, with smaller standard errors indicating better estimates.[114] In our example of the Nixon papers, the standard error for the sample average can be computed from (1) the size of the sample—500 boxes—and (2) the standard deviation of the sample values. Bigger samples give estimates that are more precise. Accordingly, the standard error should go down as the sample size grows, although the rate of improvement slows as the sample gets bigger. ("Sample size" and "the size of the sample" just mean the number of items in the sample; the "sample average" is the average value of the items in the sample.) The standard deviation of the sample comes into play by measuring heterogeneity. The less heterogeneity in the values, the smaller the standard error. For example, if all the values were about the same, a tiny sample would give an accurate estimate. Conversely, if the values are quite different from one another, a larger sample would be needed.

With a random sample of 500 boxes and a standard deviation of $2,200, the standard error for the sample average is estimated to be about $100.[115] The plaintiff's total demand was figured as the number of boxes (20,000) times the sample average ($2,000), or $40,000,000. Therefore, the standard error for the total

114. We distinguish between (1) the standard deviation of the sample data, which measures the spread in the sample values, and (2) the standard error of the sample average, which measures the likely size of the random error in the sample average. Generally, the standard error of an estimator (such as the sample average) is the standard deviation of that estimator as applied to (hypothetically) repeated samples. Courts typically use the broader term "standard deviation" when referring to the standard error. *E.g.*, Castaneda v. Partida, 430 U.S. 482 (1977).

115. The standard error for the sample average equals

$$\sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}.$$

Freedman et al., *supra* note 14, at 367–70. N stands for the size of the population, which is 20,000; n stands for the size of the sample, which is 500. The first factor, with the N's in it, is the finite sample correction factor. Here, as in many other such examples, the correction factor is so close to 1 that it can safely be ignored. (This is why the size of population usually has no bearing on the precision of the sample average as an estimator for the population average.) Next, σ is the population standard deviation. Its value is unknown but can be estimated by the sample standard deviation of $2,200. The standard error for the sample mean is therefore estimated from the data as $2,200/$\sqrt{500}$, which is nearly $100.

demand is 20,000 times the standard error for the sample average: 20,000 × $100 = $2,000,000.[116]

How is the standard error to be interpreted? Just by the luck of the draw, a few too many high-value boxes may have come into the sample, in which case the estimate of $40,000,000 is too high. Or, a few too many low-value boxes may have been drawn, in which case the estimate is too low. This is random error. The net effect of random error is unknown, because data are available only on the sample, not on the full population. However, the net effect is likely to be something close to the standard error of $2,000,000. Random error throws the estimate off, one way or the other, by something close to the standard error. The role of the standard error is to gauge the likely size of the random error.

The plaintiff's argument may be open to a variety of objections, particularly regarding appraisal methods. However, the sampling plan is sound, as is the extrapolation from the sample to the population. And there is little need for a larger sample: Relative to the total claim of $40 million, a standard error of $2 million is reasonably small.

What Is the Confidence Interval?

Although random errors larger in magnitude than the standard error are commonplace, random errors larger in magnitude than two or three times the standard error are unusual. Confidence intervals make these ideas more precise. Usually, a confidence interval for the population average (the average of all the values in the population) is centered at the sample average; the desired confidence level is obtained by adding and subtracting a suitable multiple of the standard error. In dealing with large samples, statisticians who say that the population average falls within 1 standard error of the sample average will be correct about 68% of the time. Those who say "within 2 standard errors" will be correct about 95% of the time, and those who say "within 3 standard errors" will be correct about 99.7% of the time, and so forth.

The normal curve and large samples

These confidence levels correspond to areas under a famous bell-shaped curve—the normal curve. According to a fundamental theorem of statistics (the central

116. We are assuming a simple random sample. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. In fact, the Nixon appraisers used more elaborate statistical procedures. Moreover, they valued the material as of 1995, extrapolated backward to the time of taking (1974), and then added interest. After 20 years of litigation, Nixon's estate settled with the government for $18 million. U.S. Dep't of Just., Press Release, June 12, 2000, https://perma.cc/PJC9-47CY.

limit theorem), if we were to draw not merely a single large sample from a much larger population, or even 500 of them as in the Nixon example, but millions upon millions of them (replacing the items from each sample before drawing the next one), and if we then sorted the sample averages into appropriate bins of a histogram, the heights of the bins would be greatest near the population average, and they would fall off symmetrically on each side as prescribed by the values of the normal curve centered at this point and having a standard deviation that is the standard error.[117]

The confidence interval relies on this understanding of the distribution of sample means but goes in the other direction. Knowing how the sample statistic behaves as a function of the population parameters, we draw an interval around the sample statistic that we hope will cover the true value (the parameter). The mathematical properties of the normal distribution justify the numbers used to express the "confidence."[118] In particular,

- To get a 68% confidence interval, start at the sample average, then add and subtract 1 standard error.
- To get a 95% confidence interval, start at the sample average, then add and subtract twice the standard error.
- To get a 99.7% confidence interval, start at the sample average, then add and subtract three times the standard error.

With the Nixon papers, the 68% confidence interval for plaintiff's total demand runs

- from \$40,000,000 – \$2,000,000 = \$38,000,000
- to \$40,000,000 + \$2,000,000 = \$42,000,000.

117. The normal curve is the density of a normal distribution. The distribution has two parameters—the population mean (often denoted μ) and the standard deviation (σ), which is the standard error of the variable x (here, the sample mean, which varies from one sample to the next). The equation is

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $e = 2.71828\ldots$ and $\pi = 3.14159\ldots$. Once the two parameters μ and σ are given, the density is completely specified.

118. The area under the normal curve between $x = \mu-\sigma$ and $x = \mu+\sigma$ is close to 68.3%. Likewise, the area between $\mu-2\sigma$ and $\mu+2\sigma$ is close to 95.4%. Many academic statisticians would use ±1.96 SE for a 95% confidence interval. However, the normal curve only gives an approximation to the relevant chances, and the error in that approximation will often be larger than a few tenths of a percent. For simplicity, we use ±1 SE for the 68% confidence level, and ±2 SE for 95% confidence.

The 95% confidence interval runs

- from \$40,000,000 – (2 × \$2,000,000) = \$36,000,000
- to \$40,000,000 + (2 × \$2,000,000) = \$44,000,000.

The 99.7% confidence interval runs

- from \$40,000,000 – (3 × \$2,000,000) = \$34,000,000
- to \$40,000,000 + (3 × \$2,000,000) = \$46,000,000.

To write this more compactly, we abbreviate standard error as SE. Thus, 1 SE is one standard error, 2 SE is twice the standard error, and so forth. With a large sample and an estimate like the sample average, a 68% confidence interval ranges from

estimate – 1 SE to estimate + 1 SE.

A 95% confidence interval is ranges from

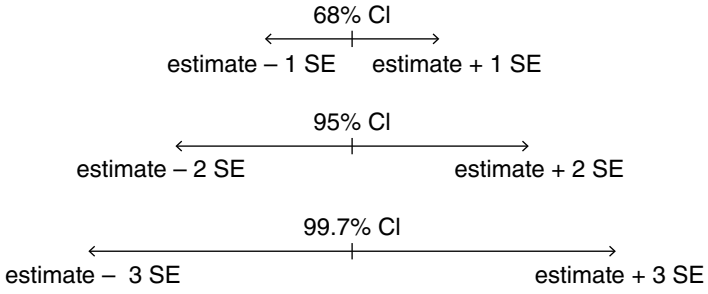
estimate – 2 SE to estimate + 2 SE.

A 99.7% confidence interval ranges from

estimate – 3 SE to estimate + 3 SE.

For a given sample size, increased confidence can be attained only by widening the interval. The 95% confidence level is the most popular, but some authors use 99%, and 90% is seen on occasion. (The corresponding multipliers on the SE are about 2, 2.6, and 1.6, respectively.) The phrase "margin of error" generally means twice the standard error. In medical journals, "confidence interval" is often abbreviated as "CI." The relationship between width and confidence is shown in Figure 4.

Figure 4. Confidence coefficients for CIs of ±1, 2, and 3 standard errors of a normally distributed estimator.



The picture shows a tradeoff between precision (the width of the interval) and confidence (that the interval covers the actual population value). In sum, an estimate based on a sample will differ from the exact population value, because of random error. The standard error gives the likely size of the random error. If the standard error is small, random error probably has little effect. If the standard error is large, the estimate may be far away from the population value. Confidence intervals are a technical refinement that provides a formalism for assessing the impact of random error or chance variation on the estimate.

Other situations

Intervals based on the standard error, with confidence levels read off the normal curve, are appropriate for estimators that are essentially unbiased and obey the central limit theorem. They generally work for sums, averages, and rates, although much depends on the design of the sampling and other matters. CIs determined via the normal curve may not work well as estimates of extremely small quantities. Table 1 of the section titled "Is the Measurement Process Valid?" presents an extreme example. The table showed that toolmark examiners given 30 pairs of cartridges from ammunition fired from the same guns and 45 pairs from different guns correctly classified every pair according to the true source. But what if the same examiners had been brought back for a second experiment with the same cartridge cases (and with no memory of the first experiment)? Would they have done as well? What about a third experiment? It would be extraordinary if they never erred in one experiment after another *ad infinitum*. Perhaps we can think of the one experiment as if it were a random sample from an infinite population of repeated experiments. Given this framework, we might want a confidence interval that will estimate the proportion of false-positive errors on the part of the examiners in an infinite stream of experiments with guns and ammunition exactly like those in the study. The observed proportion of errors in the one experiment is zero, which makes the plug-in estimate of the standard error zero as well. However, a confidence interval of zero width cannot be right. The probability of false-positive errors could be higher than zero, and an examiner could still get all 45 determinations for the false pairs correct. The problem is that because every determination in the sample (the one experiment) is correct, there is no variability to inform what we might see in future samples (experiments).

How should we proceed? It is easy to show that an interval from 0 to 6.5% would achieve at least 95% "confidence."[119] So if we are willing to embrace anything in the range of 0 to 6.5% as the true error-rate when given samples with an

119. When the long-run false-positive error rate is 6.5%, the probability of 45 independent correct classifications of cartridges from different pairs of guns is $(1-0.065)^{45} = 0.049$. When the "true" error rate is greater than 6.5%, that probability for zero errors in the sample is smaller still.

observed rate of 0/45, we will be correct at least 95% of the time (in the long run). But this interval is a very conservative estimate. How to construct a confidence or other interval for the probability of an event when 0/*n* are observed goes by the name of the zero-numerator problem, and several approaches have been proposed.[120]

In still other situations (for example, where observations are dependent), other methods can be used to obtain confidence intervals. One class of methods repeatedly resamples from the observed data to approximate what would happen in repeated samples from the full population.[121] With modern computers, the resulting bootstrap confidence intervals are enticingly easy to calculate, but they too require certain assumptions to be verified.[122]

How Big Should the Sample Be?

There is no easy answer to this sensible question. Much depends on the level of error that is tolerable, the material being sampled, and the sampling method. Increasing the size of the sample provides no protection against bias ("nonsampling error"). Indeed, beyond some point, large samples are harder to manage and more vulnerable to nonsampling error. To reduce bias, the researcher must improve the design of the study or use a statistical model more tightly linked to the data-collection process. Larger samples generally will reduce the level of random error ("sampling error"). It rarely will be sensible to draw a probability sample with fewer than, say, two or three dozen items, and with such small samples, methods based solely on the normal curve (see section titled "What Is the Standard Error?" above) will not apply. A pilot sample is often valuable, partly to obtain an initial estimate of characteristics of the population that may be used in

120. *E.g.*, Noriah M. Al-Kandari & Paul H. Garthwaite, *Bayesian Analysis of Misclassified Binomial Data: Double-Sampling and the Zero-Numerator Problem*, Communications in Statistics—Simulation & Computation (2020), https://doi.org/10.1080/03610918.2020.1855448; B. D. Jovanovic & P. S. Levy, *A Look at the Rule of Three*, 51 Am. Statistician 137 (1997); Robert L. Winkler et al., *The Role of Informative Priors in Zero-Numerator Problems: Being Conservative Versus Being Candid*, 56 Am. Statistician 1 (2002).

121. *See, e.g.,* Freedman, *supra* 25, at 147–50; David H. Kaye, *Frequentist Statistical Inference*, *in* Handbook of Forensic Statistics 39, 66–68 (David Banks et al. eds., 2021) (discussing the basic idea). On the patentability of procedures relying on bootstrapping or other resampling methods, see *SAP America v. InvestPic*, 898 F.3d 1161 (Fed. Cir. 2018) (investment analysis); *Cybergenetics Corp. v. Inst. of Env'l Sci. & Rsch.*, 490 F. Supp. 3d 1237 (N.D. Ohio 2020) (in discerning separate contributors to DNA mixtures), *aff'd*, 856 Fed. App'x 312 (Fed. Cir. 2021).

122. *See In re* Sonic Corp. Customer Data Sec. Breach Litig., No. 1:17-md-2807, MDL No. 2807, 2021 WL 5916743 (N.D. Ohio Dec. 15, 2021) (bootstrap estimate that losses to credit-card companies from a computer security breach exceeded $76,000,000 were inadmissible because the bootstrap sample was small and unrepresentative); Bradley Efron & Robert J. Tibshirani, An Introduction to the Bootstrap (1994); Freedman, *supra* note 25, at 150.

determining the final sample size.[123] If a population appears to be heterogeneous, then alternatives to simple random sampling, such as stratified random samples (that are each fairly homogeneous), may provide more representative samples and more precise estimates for the same total sample size. As these examples illustrate, probability samples require some effort in the design phase.

Population size (i.e., the number of items in the population) usually has little bearing on the precision of estimates for the population average. This is surprising. On the other hand, population size has a direct bearing on estimated totals. Both points are illustrated by the Nixon papers (see section titled "What Is the Standard Error?" above). To be sure, drawing a probability sample from a large population may involve a lot of work. Samples presented in the courtroom have ranged from 5 (tiny) to 1.7 million (huge).[124]

What Are the Technical and Interpretive Difficulties with Confidence Intervals?

To begin with, "confidence" has an esoteric meaning in this context. The confidence level indicates the percentage of the time that intervals from repeated samples would cover the true value. The confidence level does not express the chance that repeated estimates would fall into the stated confidence interval.[125]

123. Suppose a researcher is interested in the association between alcohol intoxication and single-vehicle accidents. A small sample of accident records (along with intuitions) could suggest some guess for the population proportion of single-vehicle crashes in which the driver was found to be intoxicated. Assume the guess is 50%. Suppose further that the researcher is willing to tolerate errors of up to ±3% in the estimate from the larger sample being planned. Finally, suppose that the researcher plans to report a conventional 95% confidence interval. It can be shown that a simple random sample of approximately 1,067 single-vehicle accident records should suffice. Hence, the researcher collects 1,067 records at random and determines the proportion in which intoxication was in the accident record. The observed proportion in this sample might be 40% or 0.4, to pick a concrete number. The earlier guess of 50% no longer matters—it was just used to get a sense of the sample size for the full study, and the researcher now can report a 95% CI from the full study. The estimate becomes $.40 \pm 1.96\sqrt{(.40)(.60)/1067} = .40 \pm .03$, which is about 37% to 43%.

124. Lebrilla v. Farmers Grp., Inc., No. 00-CC-017185 (Cal. Super. Ct., Orange Cnty., Dec. 5, 2006) (preliminary approval of settlement). This was a class action lawsuit on behalf of plaintiffs who were insured by Farmers and had automobile accidents. Plaintiffs alleged that replacement parts recommended by Farmers did not meet specifications: Small samples were used to evaluate these allegations. At the other extreme, it was proposed to adjust Census 2000 for undercount and overcount by reviewing a sample of 1.7 million persons. *See* Brown et al., *supra* note 36, at 353.

125. Opinions reflecting this misinterpretation include *Turpin v. Merrell Dow Pharm., Inc.*, 959 F.2d 1349, 1353 (6th Cir. 1992) ("If a confidence interval of '95 percent between 0.8 and 3.10' is cited, this means that random repetition of the study should produce, 95 percent of the time, a relative risk somewhere between 0.8 and 3.10."); *Garcia v. Tyson Foods, Inc.*, 890 F. Supp. 2d 1273, 1285 (D. Kan. 2012) ("Dr. Radwin testified that his study was conducted within a confidence interval of 95—that is 'if I did this study over and over again, 95 out of a hundred times I would

With the Nixon papers, the 95% confidence interval should not be interpreted as saying that 95% of all random samples will produce estimates in the range from $36 million to $44 million. Rather, a high degree of "confidence" makes it reasonable to accept the one sample interval as a plausible statement of what the value of all the papers could be.

Second, the confidence level does not give the probability that the unknown parameter lies within the confidence interval.[126] For example, the 95% confidence level should not be translated to a 95% probability that the total value of the papers is in the range from $36 million to $44 million. According to the frequentist theory of statistics, probability statements cannot be made about population characteristics: Probability statements apply to the behavior of samples. That is why the different term "confidence" is used.

Third, this "confidence" is a statement about the entire region. One might think that values near the center of the interval are more likely to be closer to the true value than those at the extremes. Yet the statement of confidence applies equally to all the values in the interval. Moreover, that the intervals have sharp boundaries does not imply that there is an important difference between a value at the edge of the interval and one just beyond it.

expect to get an average between that interval.'"); *In re Silicone Gel Breast Implants Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 897 (C.D. Cal. 2004) ("a margin of error between 0.5 and 8.0 at the 95% confidence level . . . means that 95 times out of 100 a study of that type would yield a relative risk value somewhere between 0.5 and 8.0").

Language from another reference guide in the previous edition of this *Reference Manual* that is often quoted may inadvertently convey the incorrect impression that a confidence coefficient such as 95% refers to the percentage of results in (hypothetically) repeated studies that would be expected to lie within the interval reported in the study before the court. *See, e.g.,* Rhyne v. U.S. Steel Corp., 474 F. Supp. 3d 733, 744 (W.D.N.C. 2020) ("'If a 95% confidence interval is specified, the range encompasses the results we would expect 95% of the time if samples for new studies were repeatedly drawn from the population.' Reference Guide on Epidemiology, at 580."). However, simulations suggest that the confidence coefficient for a sample mean from a normal population tends to overstate the probability that the next sample mean will lie within the interval. Geoff Cumming & Robert Maillardet, *Confidence Intervals and Replication: Where Will the Next Mean Fall?*, 11 Psych. Methods 217 (2006), https://doi.org/10.1037/1082-989X.11.3.217 (finding that "[o]n average, a 95% CI will include just 83.4% of future replication means"). The more technically correct statement in the *Silicone Gel* case would be that "the confidence interval of 0.5 to 8.0 means that the relative risk in the population plausibly could fall within this wide range—and that in roughly 95 times out of 100 random samples from the same population, the confidence intervals (however wide they might be) would include the population value (whatever it is)."

126. *See, e.g.*, Freedman et al., *supra* note 14, at 383–86. Consequently, it is misleading to suggest that "[t]he confidence interval means . . . it is 95 percent likely that the actual [value of the parameter being estimated] is somewhere in the interval, somewhere between the low end of the interval and its high end," *Center for Biological Diversity v. U.S. Fish & Wildlife Serv.*, 342 F. Supp. 3d 968, 977 (N.D. Cal. 2018), or that "a 95-percent median confidence interval of 89.43 to 90.28 percent [means] that the true median is 95-percent likely to fall within that range," *Cnty. of Douglas v. Nebraska Tax Equalization & Rev. Comm'n*, 894 N.W.2d 308, 316 (Neb. 2017).

Fourth, for a given confidence level, a narrower interval indicates a more precise estimate, whereas a broader interval indicates less precision.[127] A high confidence level with a broad interval means very little, but a high confidence level for a small interval is impressive, indicating that the random error in the sample estimate is low. For example, take a 95% confidence interval for a damage claim. An interval that runs from $36 million to $44 million is more precise than a much wider interval that goes from $10 million to $70 million. Statements about confidence without mention of an interval are practically meaningless.[128]

The final point to make is that standard errors and confidence intervals are often derived from statistical models for the process that generated the data. The model usually has parameters—numerical constants describing the population from which samples were drawn. When the values of the parameters are not known, the statistician must work backwards, using the sample data to make estimates. That was the case for valuing the Nixon papers. One parameter is the average value of all 20,000 boxes, and another parameter is the standard deviation of the 20,000 values.[129] Generally, the probabilities used in drawing inferences are computed from a model with estimated parameter values.

127. *See* section titled "What Is the Standard Error?" above. In *Cimino v. Raymark Industries, Inc.*, 751 F. Supp. 649 (E.D. Tex. 1990), *rev'd*, 151 F.3d 297 (5th Cir. 1998), the district court drew certain random samples from more than 6,000 pending asbestos cases, tried these cases, and used the results to estimate the total award to be given to all plaintiffs in the pending cases. The court then held a hearing to determine whether the samples were large enough to provide accurate estimates. The court's expert, an educational psychologist, testified that the estimates were accurate because the samples matched the population on such characteristics as race and the percentage of plaintiffs still alive. *Id.* at 664. However, the matches occurred only in the sense that population characteristics fell within 99% confidence intervals computed from the samples. The court thought that matches within the 99% confidence intervals proved more than matches within 95% intervals. *Id.* This is backward. To be correct in a few instances with a 99% confidence interval is not very impressive—by definition, such intervals are broad enough to ensure coverage approximately 99% of the time.

128. In *Hilao v. Estate of Marcos*, 103 F.3d 767 (9th Cir. 1996), "an expert on statistics . . . testified that . . . a random sample of 137 claims would achieve 'a 95% statistical probability that the same percentage determined to be valid among the examined claims would be applicable to the totality of [9,541 facially valid] claims filed.'" *Id.* at 782. There is no 95% "statistical probability" that a percentage computed from a sample will be "applicable" to a population. One can compute a confidence interval from a random sample and be 95% confident that the interval covers some parameter. The computation can be done for a sample of virtually any size, with larger samples giving smaller intervals. What is missing from the opinion is a discussion of the widths of the relevant intervals. For the same reason, it is meaningless to testify, as an expert did in *Ayyad v. Sprint Spectrum, L.P.*, No. RG03–121510 (Cal. Super. Ct., Alameda Cnty.) (transcript, May 28, 2008, at 730), that a simple regression equation is trustworthy because the coefficient of the explanatory variable has "an extremely high indication of reliability to more than 99% confidence level."

129. These parameters can be used to approximate the distribution of the sample average. *See* section titled "What Is the Standard Error?" above. Regression models and their parameters are discussed in the section titled "Correlation and Regression" below and in Rubinfeld & Card, *supra* note 25.

If the data come from a probability sample or a randomized controlled experiment (see sections titled "Descriptive Surveys and Censuses" and "Is the Study Designed to Investigate Causation?" above), then the statistical model may be connected tightly to the actual data-collection process. In other situations, using the model may be tantamount to assuming that a sample of convenience is like a random sample, or that an observational study is like a randomized experiment. With the Nixon papers, the appraisers drew a random sample, and that justified the statistical calculations—if not the appraised values themselves. In many contexts, the choice of an appropriate statistical model is less than obvious. When a model does not fit the data-collection process, estimates and standard errors will not be probative.

Standard errors and confidence intervals are designed to take account of random errors but not systematic ones such as selection bias or nonresponse bias (see sections titled "What Method Is Used to Select the Units?" and "Of the Units Selected, Which Provide Measurements?" above). For example, after reviewing studies to see whether a particular drug caused birth defects, a court observed that mothers of children with birth defects may be more likely to remember taking a drug during pregnancy than mothers with normal children.[130] This selective recall would bias comparisons between samples from the two groups of women. Neither the standard error nor the confidence interval for the difference in drug usage between the groups accounts for this bias.[131]

p-values, Significance Levels, and Hypothesis Tests

What Is the p-value?

In 1969, Dr. Benjamin Spock came to trial in the U.S. District Court for Massachusetts. The charge was conspiracy to violate the Military Service Act. The jury was drawn from a panel of 350 persons selected by the clerk of the court. The panel included only 102 women—substantially less than 50%—although a majority of the eligible jurors in the community were female. The shortfall in women was especially poignant in this case: "Of all defendants, Dr. Spock, who had given

130. Brock v. Merrell Dow Pharm., Inc., 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989).

131. In *Brock*, the court stated that the confidence interval took account of bias (in the form of selective recall) as well as random error. 874 F.2d at 311–12. This is wrong. Even if the sampling error were nonexistent—which would be the case if one could interview every woman who had a child during the period that the drug was available—selective recall would produce a difference in the percentages of reported drug exposure between mothers of children with birth defects and those with normal children. In this hypothetical situation, the standard error would vanish. Therefore, the standard error could disclose nothing about the impact of selective recall.

wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury."[132]

Can the shortfall in women be explained by the mere play of random chance? To approach the problem, a statistician could formulate and test a null hypothesis. Here, the null hypothesis says that the panel is like 350 persons drawn at random from a large population that is 50% female. The expected number of women drawn would then be 50% of 350, which is 175. The observed number of women is 102. The shortfall is 175 – 102 = 73. How likely is it to find a disparity this large or larger, between observed and expected values? The probability is called the *p*-value, or *p*.

The *p*-value is the probability of getting data as extreme as, or more extreme than, the actual data—given that the null hypothesis is true. In the example, *p* can be computed from a simple probability model, expressed as a function that gives the probability of every possible outcome for the number of women in a sample. A reasonable model here, known as the binomial distribution, depends on just two parameters: the size n of the sample and a constant probability θ of selecting a woman on each draw.[133] Using the binomial formula, the probability of a sample with a shortfall or an excess of at least 73 women turns out to be essentially zero.[134] The discrepancy between the observed and the expected is far too large to explain by random chance. Indeed, even if the panel had included 155 women, the *p*-value would only be around 0.04, or 4%.[135] (If the population is more than 50% female, *p* will be even smaller.) In short, the jury panel was nothing like a random sample from the community.

Large *p*-values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance

132. Hans Zeisel, *Dr. Spock and the Case of the Vanishing Women Jurors*, 37 U. Chi. L. Rev. 1 (1969). Zeisel's reasoning was different from that presented in this text. The conviction was reversed on appeal without reaching the issue of jury selection. United States v. Spock, 416 F.2d 165 (1st Cir. 1965).

133. In mathematical notation, the probability for the number x of women in a sample of size n is the function $f(x;\ n,\ \theta) = n!/[x!(n-x)!] \times \theta^n(1-\theta)^n$. This binomial formula for f is discussed in many texts, including Freedman et al., *supra* note 14, at 255–61. This model is a good choice when the population of potential jurors is so large that the 50–50 split ($\theta = 0.5$) of men and women does not change appreciably as each juror is selected. A different distribution (called a hypergeometric distribution) would be used if that complication were more important. *See, e.g.*, David H. Kaye, *Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit*, 26 Jurimetrics J. 215 (1986).

134. The binomial probability of getting exactly 102 women in the sample is $f(x = 102;\ n = 350,\ \theta = ½)$, which works out to $1/10^{15}$. With the binomial probability function, we can compute the chance of getting 101 women, or 100, or any other particular number x. The chance of getting 102 women or fewer is then computed by addition. The chance is about $2/10^{15}$. The number 10^{15} is 1 followed by 15 zeros, that is, a quadrillion. The chance of an equally extreme outcome in the other direction ($x = 248$ or more) is the same. The *p*-value of $4/10^{15}$ is very bad news for the null hypothesis.

135. *See* Kaye & Freedman, *supra* note 111, at Appendix B.

variation. On the other hand, if p is very small, something other than chance must be involved: The data are far away from the values expected under the null hypothesis. Significance testing often seems to involve multiple negatives. This is because a statistical test is an argument by contradiction. With the Dr. Spock example, the null hypothesis asserts that the jury panel is like a random sample from a population that is 50% female. The data contradict this null hypothesis because the disparity between what is observed and what is expected (according to the null) is too large to be explained as the product of random chance. In a typical jury discrimination case, small p-values help a defendant appealing a conviction by showing that the jury panel is not like a random sample from the relevant population; large p-values hurt the defendant's case. Likewise, in the usual employment context, small p-values help plaintiffs who complain of discrimination—for example, by showing that a disparity in promotion rates is too large to be explained by chance; conversely, large p-values would be consistent with the defense argument that the disparity is just due to chance.

Because p is calculated by assuming that the null hypothesis is correct, p does not give the chance that the null is true. The p-value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the p-value can therefore be summarized in two lines:

> p is the probability of extreme data given the null hypothesis.
> p is not the probability of the null hypothesis given extreme data.[136]

136. Some opinions present a contrary view. *E.g.*, Berghuis v. Smith, 559 U.S. 314, 324 n.1 (2010) (noting that statistical analysis "seeks to determine the probability that the disparity between a group's jury-eligible population and the group's percentage in the qualified jury pool is attributable to random chance"); Vasquez v. Hillery, 474 U.S. 254, 259 n.3 (1986) ("the District Court . . . ultimately accepted . . . a probability of 2 in 1000 that the phenomenon was attributable to chance"); United States v. Carter, 750 F.3d 462, 468 n.15 (4th Cir. 2014) ("a p-value of 0.068 indicates that there was only a 6.8% chance that the correlation was due to chance"). Such statements confuse the probability of the kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome—the "transposition fallacy."

Instances of the transposition fallacy in criminal cases are collected in Kaye et al., *supra* note 8, §§ 12.8.2(b) & 14.1.2. In *McDaniel v. Brown*, 558 U.S. 120 (2010), for example, a DNA analyst suggested that a random-match probability of 1/3,000,000 implied a .000033 probability that the defendant was not the source of the DNA found on the victim's clothing. *See* David H. Kaye, *The Interpretation of DNA Evidence: A Case Study in Probabilities, in* Science Policy Decision-Making Educational Modules (Nat'l Acad. Sci., Eng'g & Med. Comm. on Preparing the Next Generation of Policy Makers for Science-Based Decisions ed., 2016), https://www.nationalacademies.org/our-work/science-policy-decision-making-educational-modules.

To recapitulate the logic of *p*-values: If *p* is small, the observed data are far from what is expected under the null hypothesis—too far to be readily explained by the operations of chance. That discredits the null hypothesis.

Computing *p*-values requires statistical expertise. Many methods are available, but only some will fit the occasion. Sometimes standard errors will be part of the analysis, other times they will not be.[137] Sometimes a difference of two standard errors will imply a *p*-value of about 5%; other times it will not. In general, the *p*-value depends on the statistical model, the size of the sample, and the sample statistics.

Is a Difference Statistically Significant?

If an observed difference is in the middle of the distribution that would be expected under the null hypothesis, there is no surprise. The sample data are of the type that often would be seen when the null hypothesis is true. The observed difference (or similar statistic) does not fall into a region that is appropriate for rejecting the null hypothesis. It is not "significant," as statisticians are wont to say. On the other hand, if the sample difference is far from the expected value—according to the null hypothesis—then the sample is unusual. The difference is significant, and the null hypothesis is rejected.

Statistical significance can be determined by comparing *p* to a preset value, called the significance level.[138] In this approach, the null hypothesis is rejected when *p* falls below this level. In practice, statistical analysts typically use levels of

137. The binomial analysis in the *Spock* case did not use standard errors. The standard error there is the square root of $n\theta(1-\theta) = 9.35$. The observed value of 102 is nearly 8 standard errors below the expected value of 175, which is a lot of standard errors. However, we did not have to perform this "standard deviation analysis" (Berghuis v. Smith, 559 U.S. 314, 324 n.1 (2010); *infra* note 139) to see that the observed disparity was incompatible with the null hypothesis. The *p*-value told us that directly. Using a particular number of standard errors to mark statistical significance comes from the tradition of using the normal curve to approximate the distribution of sample statistics such as the sample proportion or mean. *See* section titled "What Is the Standard Error?" above.

138. It is not necessary to compute the *p*-value to perform the test for significance. Instead, taking the Neyman-Pearson approach mentioned in note 105, before analyzing the data, one can define a rejection region that keeps the risk of falsely rejecting the null hypothesis at or below a prespecified level (and that maximizes the power of the test to detect a difference if one is present). Statisticians use the Greek letter alpha (α) to denote this level; α gives the chance of getting a result that produces a rejection, assuming that the null hypothesis is true. Thus, α represents the chance of a false rejection of the null hypothesis (also called a false positive, a false alarm, or a Type I error). For example, suppose $\alpha = 5\%$. If investigators do many studies, and the null hypothesis happens to be true in each case, then about 5% of the time they would obtain significant results—and falsely reject the null hypothesis. Inasmuch as the data in the rejection region are all such that $p \leq \alpha$, it is common to describe the test, as we have in the text, in terms of the *p*-value and to call $\alpha = 0.05$ the significance level.

5% and 1%.[139] The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure. An unexplained reference to highly significant results probably means that p is less than 1%. These levels of 5% and 1% have become icons of science and the legal process. In truth, however, such levels are simply common conventions.[140]

Because the term "significant" is merely a label for a certain kind of p-value, significance is subject to the same limitations as the underlying p-value. Thus, significant differences may be evidence that something besides random error is at work. They are not evidence that this something is legally or practically important. Statisticians distinguish between statistical and practical significance to make the point. When practical significance is lacking—when the size of a disparity is negligible[141]—statistical significance may be of no legal significance.[142]

139. The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as "suspect to a social scientist" when a statistic from "large samples" falls more than "two or three standard deviations" from its expected value under the null hypothesis. Although the Court did not say so, these differences produce p-values of about 5% and 0.3% when the statistic is normally distributed. The Court's standard deviation is our standard error.

140. Some opinions quote statisticians who urge giving readers p-values instead of making pronouncements of "significant" or "not significant" or who "caution against using statistical significance rigidly" as if these authorities are recommending admission of any relevant and properly executed study, regardless of the p-value. *E.g.*, *In re* Urethane Antitrust Litig., 166 F. Supp. 3d 501, 508–09 (D.N.J. 2016) (relying on such statements to justify admission of findings with p-values of 7.2%, 19.2%, and, in dictum, 50%). However, declarations that p-values are more informative than pronouncements of "significance" do not mean that unimpressive results (those with large p-values) are admissible as proof of the alternative hypothesis. Large p-values indicate that the observed results are not reliable evidence for the alternative hypothesis; consequently, under Rules 403 and 702, very large p-values go to the admissibility as well as the weight of the statistical evidence.

141. Context affects judgments of what outcomes lack practical significance. Some relatively small differences can have a large practical effect. For example, a loss of half a percentage point in the revenue of a corporation that is a consequence of a competitor's illegal conduct would be practically significant in computing damages if the total revenue was large.

142. *E.g.*, Brnovich v. Democratic Nat'l Comm., 141 S. Ct. 2321, 2343 n.17 (2021) (quoting substantially identical text in the third edition of this Manual); *id.* at 2358 n.4 (dissenting opinion agreeing that "there may be some threshold of what is sometimes called 'practical significance'—a level of inequality that, even if statistically meaningful, is just too trivial for the legal system to care about") (dissenting opinion); Waisome v. Port Auth., 948 F.2d 1370, 1376 (2d Cir. 1991) ("though the disparity was found to be statistically significant, it was of limited magnitude"); United States v. Hernandez-Estrada, 749 F.3d 1154, 1165 (9th Cir. 2014) (en banc) ("[T]he challenging party must establish not only statistical significance, but also legal significance. . . . In other words, if a statistical analysis shows underrepresentation, but the underrepresentation does not substantially affect the representation of the group in the actual jury pool, then the underrepresentation does not have legal significance in the fair cross-section context."); Apsley v. Boeing Co., 691 F.3d 1184 (10th Cir. 2012) (a reported p-value of 1/50,000 with respect to hiring older workers was not sufficient to defeat defendant's motion for summary judgment when the difference between the expected and

It is easy to mistake the *p*-value for the probability of the null hypothesis given the data (see section titled "What Is the *p*-value?" above). Likewise, if results are significant at the 5% level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.[143] This temptation should be resisted. From the frequentist perspective, statistical hypotheses are either true or false. Probabilities govern the samples, not the models and hypotheses. The significance level tells us what is likely to happen when the null hypothesis is correct; it does not tell us the probability that the hypothesis is true. Significance comes no closer to expressing the probability that the null hypothesis is true than does the underlying *p*-value.

Recent Emphasis on the Limitations of p-values

For many of the reasons mentioned above, there has been considerable controversy about issues related to *p*-values. In response to the perception that many published studies reporting significance with the .05 criterion are not reproducible—the much bruited "replication crisis"[144] in psychology, medicine, and other

actual number hired was only about 50 out of nearly 7,300); United States v. Henderson, 409 F.3d 1293, 1306 (11th Cir. 2005) (regardless of statistical significance, excluding law enforcement officers from jury service does not have a large enough impact on the composition of grand juries to violate the Jury Selection and Service Act); *cf.* Thornburg v. Gingles, 478 U.S. 30, 53–54 (1986) (repeating the district court's explanation of why "the correlation between the race of the voter and the voter's choice of certain candidates was [not only] statistically significant," but also "so marked as to be substantively significant, in the sense that the results of the individual election would have been different depending upon whether it had been held among only the white voters or only the black voters"); cases cited, *infra* notes 149–150. *But see* Jones v. City of Boston, 752 F.3d 38, 53 (1st Cir. 2014) (in Title VII cases, "a plaintiff's failure to demonstrate practical significance cannot preclude that plaintiff from relying on competent evidence of statistical significance to establish a prima facie case of disparate impact").

143. *E.g.*, *Waisome*, 948 F.2d at 1376 ("Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random. . . ."); Adams v. Ameritech Serv., Inc., 231 F.3d 414, 424 (7th Cir. 2000) ("Two standard deviations is normally enough to show that it is extremely unlikely (. . . less than a 5% probability) that the disparity is due to chance."); Magistrini v. One Hour Martinizing Dry Cleaning, 180 F. Supp. 2d 584, 605 n.26 (D.N.J. 2002) (a "statistically significant . . . study shows that there is only 5% probability that an observed association is due to chance"); *cf.* Giles v. Wyeth, Inc., 500 F. Supp. 2d 1048, 1056 (S.D. Ill. 2007) ("While [plaintiff] admits that a *p*-value of .15 is three times higher than what scientists generally consider statistically significant—that is, a *p*-value of .05 or lower—she maintains that this 'represents 85% certainty, which meets any conceivable concept of preponderance of the evidence.'").

144. John P.A. Ioannidis, *Why Most Clinical Research Is Not Useful*, 13 PLOS Med. E1002049 (2016), https://doi.org/10.1371/journal.pmed.1002049; Patrick E. Shrout & Joseph L. Rodgers, *Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis*, 69 Ann. Rev. Psych. 487 (2018), https://doi.org/10.1146/annurev-psych-122216-011845. For a description in the legal literature, see Beerdsen, *supra* note 10.

fields—prominent manifestos to "retire statistical significance"[145] and to move to "a world beyond 'p < 0.05'" have appeared.[146] Various alternatives or supplements to *p*-values are available, and the American Statistical Association (ASA) has issued statements on the best use of *p*-values and other quantities.[147] Its 2016 statement emphasized that *p*-values can be used to indicate the degree to which data are incompatible with a specified statistical model, but that they are not the probability that the model is true; that the size of the difference rather than just statistical significance should be reported; and that scientific conclusions or business decisions should not be based only on whether a test yields a statistically significant result.[148]

Tests or Interval Estimates?

How can a highly significant difference be practically insignificant? The reason is simple: *p* depends not only on the magnitude of the effect, but also on the

145. Valentin Amrhein et al., *Retire Statistical Significance*, 567 Nature 305 (2019), https://doi.org/10.1038/d41586-019-00857-9 (statement endorsed by more than 200 statisticians and other scientists calling "for a stop to the use of P values in the conventional, dichotomous way—to decide whether a result refutes or supports a scientific hypothesis. . . . . P values, intervals and other statistical measures all have their place, but it's time for statistical significance to go.").

146. In a special issue of *The American Statistician*, the executive director of the American Statistical Association (ASA) and other authors proclaimed that "it is time to stop using the term 'statistically significant' entirely." Ronald L. Wasserstein et al., *Moving to a World Beyond "p < 0.05,"* 73:sup1 Am. Statistician 1, 2 (2019), https://doi.org/10.1080/0031305.2019.1583913 (adding that "[n]or should variants such as 'significantly different,' 'p < 0.05,' and 'nonsignificant' survive, whether expressed in words, by asterisks in a table, or in some other way"). However, the recommendation was not "official ASA policy." Karen Kafadar, Editorial, *Statistical Significance, P-values, and Replicability*, 15 Annals Applied Stat. 1081 (2021), https://doi.org/10.1214/21-AOAS1500. For context on the *Nature* and ASA statements, see Benjamini, *supra* note 10.

147. Surprisingly, when $p = 0.05$, it is not very probable that a repetition of the same study under identical conditions would succeed in producing results for which $p < 0.05$. *E.g.*, Leonhard Held et al., *Replication Power and Regression to the Mean*, Significance, Dec. 2020, at 10–11 (arguing that this shows "the need for more stringent p-value thresholds to trust (original) 'out-of-the-blue' findings"); Laura C. Lazzeroni et al., *Solutions for Quantifying P-value Uncertainty and Replication Power*, 13 Nature Methods 107 (2016), https://doi.org/10.1038/nmeth.3741.

148. Ronald L. Wasserstein & Nicole A. Lazar, *ASA Statement on Statistical Significance and P-values*, 70 Am. Statistician 129 (2016), https://dx.doi.org/10.1080/00031305.2016.1154108 (also noting in ¶ 3.6 that "a p-value near 0.05 taken by itself offers only weak evidence"). A task force appointed by the president of the American Statistical Association ecumenically concluded that "P-values, confidence intervals and prediction intervals [as well as] Bayes factors, posterior probability distributions and credible intervals are . . . some among many statistical methods useful for reflecting uncertainty" and that "P-values and significance tests, *when properly applied and interpreted*, increase the rigor of the conclusions drawn from data." Yoav Benjamini et al., *ASA President's Task Force Statement on Statistical Significance and Replicability*, 15 Annals Applied Stat. 1084 (2021), https://doi.org/10.1214/21-AOAS1501 (emphasis added). On Bayes factors and the like, see section titled "Bayesian Statistical Methods and Posterior Probabilities" below.

sample size (among other things). With a huge sample, even a tiny effect will be highly significant.[149] For example, suppose that a company hires 52% of male job applicants and 49% of female applicants. With a large enough sample, a statistician could compute an impressively small *p*-value. This *p*-value would confirm that the difference does not result from chance, but it would not convert a minor difference (52% versus 49%) into a substantial one.[150] In short, the *p*-value does not measure the strength or importance of an association.

A "significant" effect can be small. Conversely, an effect that is "not significant" can be large. By inquiring into the magnitude of an effect, courts can avoid being misled by *p*-values. To focus attention on more substantive concerns—the size of the effect and the precision of the statistical analysis—interval estimates (e.g., confidence intervals) may be more valuable than tests. Seeing a plausible range of values for the quantity of interest helps describe the statistical uncertainty in the estimate.

Is the Sample Statistically Significant?

Many a sample has been praised for its statistical significance or blamed for its lack thereof. Technically, this makes little sense. Statistical significance is about the difference between observations and expectations. Significance therefore applies to statistics computed from the sample, but not to the sample itself, and certainly not to the size of the sample. Findings can be statistically significant. Differences can be statistically significant (see section titled "Is a Difference Statistically Significant?" above). Estimates can be statistically significant (see section titled "Statistical Models" below). By contrast, samples can be representative or unrepresentative. They can be chosen well or badly (see section titled "What Method Is Used to Select the Units?" above). They can be large enough to give reliable results or too small to bother with (see section titled "What is the Confidence Interval?" above). But samples cannot be "statistically significant," if this technical phrase is to be used as statisticians use it.

149. *See* section titled "Is a Difference Statistically Significant?" above. Although some opinions seem to equate small *p*-values with "gross" or "substantial" disparities, most courts recognize the need to decide whether the underlying sample statistics reveal that a disparity is large. *E.g.*, Washington v. People, 186 P.3d 594 (Colo. 2008) (jury selection).

150. *Cf.* Frazier v. Garrison Indep. Sch. Dist., 980 F.2d 1514, 1526 (5th Cir. 1993) (rejecting claims of intentional discrimination in the use of a teacher competency examination that resulted in retention rates exceeding 95% for all groups); *Washington*, 186 P.2d 594 (although a jury selection practice that reduced the representation of "African-Americans [from] 7.7 percent of the population [to] 7.4 percent of the county's jury panels produced a highly statistically significant disparity, the small degree of exclusion was not constitutionally significant").

Evaluating Hypothesis Tests

What Is the Power of the Test?

The power of a statistical study needs to be considered to avoid mistaking the absence of evidence for an effect with evidence for the absence of the effect. When a *p*-value is high, findings are not significant, and the null hypothesis is not rejected. This could happen for at least two reasons: (1) the null hypothesis is true; or (2) the null is false—but, by chance, the data happened to be of the kind expected under the null. If the power of a statistical study is low, the second explanation may be plausible. Power is the chance that a statistical test will declare an effect when there is an effect to be declared.[151] This chance depends on the size of the effect and the size of the sample. Discerning subtle differences requires large samples; small samples may fail to detect substantial differences.

When a study with low power fails to show a significant effect, the results may therefore be more fairly described as inconclusive rather than negative. The proof is weak because power is low. On the other hand, when studies have a good chance of detecting a meaningful association, failure to obtain significance can be persuasive evidence that there is nothing much to be found.[152]

151. More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis (*see* section titled "What Are the Rival Hypotheses?" below) is right. Typically, this probability will depend on the values of unknown parameters, as well as the preset significance level α. The power can be computed for any value of α and any choice of parameters satisfying the alternative hypothesis. Frequentist hypothesis testing keeps the risk of a false positive to a specified level (such as $\alpha = 5\%$) and then tries to maximize power.

Statisticians usually denote power by the Greek letter beta (β). However, some authors use β to denote the probability of *accepting* the null hypothesis when the alternative hypothesis is true; this usage is fairly standard in epidemiology. Accepting the null hypothesis when the alternative holds true is a false negative (also called a Type II error, a missed signal, or a false acceptance of the null hypothesis).

The chance of a false negative may be computed from the power. Some commentators have claimed that the cutoff for significance should be chosen to equalize the chance of a false positive and a false negative, on the ground that this criterion corresponds to the more-probable-than-not burden of proof. The argument is fallacious, because $1-\alpha$ and β do not give the probabilities of the null and alternative hypotheses. *See* sections titled "What Is the *p*-value?" and "Is a Difference Statistically Significant?" above; David H. Kaye, *Hypothesis Testing in the Courtroom*, *in* Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon 331, 341–43 (Alan E. Gelfand ed., 1987).

152. Some formal procedures (meta-analysis) are available to aggregate results across studies. *See, e.g., In re* Bextra & Celebrex Marketing Sales Practices & Prod. Liab. Litig., 524 F. Supp. 2d 1166, 1174, 1184 (N.D. Cal. 2007) (holding that "[a] meta-analysis of all available published and unpublished randomized clinical trials" of certain pain-relief medicine was admissible). In principle, the power of the collective results will be greater than the power of each study. However, these procedures have their own weakness. *See, e.g.*, Richard A. Berk & David A. Freedman, *Statistical Assumptions as Empirical Commitments, in* Punishment and Social Control: Essays in Honor of Sheldon Messinger 235, 244–48 (T.G. Blomberg & S. Cohen eds., 2d ed. 2003); Michael Oakes,

What About Small Samples?

For simplicity, the examples of statistical inference discussed here (see sections titled "Estimation" and "*p*-values, Significance Levels, and Hypothesis Tests" above) were based on large samples. Small samples also can provide useful information. Indeed, when confidence intervals and *p*-values can be computed, the interpretation is the same with small samples as with large ones.[153] The concern with small samples is not that they are beyond the ken of statistical theory, but that:

1. It is hard to validate the assumptions underlying any proposed statistical approach.
2. Because approximations based on the normal curve generally cannot be used, suitable confidence intervals may be difficult to compute for parameters of interest. Likewise, *p*-values may be difficult to compute for hypotheses of interest.[154]
3. Small samples may be unreliable, with large standard errors, broad confidence intervals, and tests having low power.

One Tail or Two?

Significance testing assesses the fit of the data to the null hypothesis within a given statistical model. In assessing whether the data fit the model, there is a choice to be made about whether to use a one-tailed or two-tailed *p*-value. The terms refer to the outcomes that would be considered as evidence of a lack of fit. The issue is easily explained with an example. Suppose we toss a coin 1,000 times and get 532 heads. The null hypothesis to be tested asserts that the coin is fair. The expected number of heads is 500; the excess number of heads is 32. If the null is correct, the chance of getting 532 or more heads is 2.3%. That would be used in a

Statistical Inference: A Commentary for the Social and Behavioral Sciences (1986); Diana B. Petitti, Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine (2d ed. 2000).

153. Advocates sometimes contend that samples are "too small to allow for meaningful statistical analysis," *United States v. New York City Bd. of Educ.*, 487 F. Supp. 2d 220, 229 (E.D.N.Y. 2007), and courts often look to the size of samples from earlier cases to determine whether the sample data before them are admissible or convincing. *Id.* at 230; Timmerman v. U.S. Bank, 483 F.3d 1106, 1116 n.4 (10th Cir. 2007). However, a meaningful statistical analysis yielding a significant result can be based on a small sample, and reliability does not depend on sample size alone (see section titled "What Is the Confidence Interval?" above and the section titled "What Are the Slope and Intercept?" below). Well-known small-sample techniques include the sign test and Fisher's exact test. *E.g.*, Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers 154–56, 339–41 (2d ed. 2001); *see generally* E.L. Lehmann & H.J.M. d'Abrera, Nonparametrics (2d ed. 2006).

154. With large samples, approximate inferences (based on the central limit theorem, for example) may be quite adequate. These approximations will not be satisfactory for small samples unless the sampled values (not just the sample means) are approximately normally distributed.

one-tailed test, whose *p*-value is 2.3%. This test would be appropriate if we are only concerned that the coin is biased towards heads and thus are concerned only if we see too many heads. To make a two-tailed test, the statistician also computes the chance of a deficiency of 32 or more heads (that is, of getting 500 – 32 = 468 heads or fewer). This probability is also 2.3%. So the probability of either an excess as large or larger than what occurred *or* a comparable deficiency is 2.3% + 2.3% = 4.6%. This larger probability is the two-tailed *p*-value. This value would be appropriate if we are concerned that the coin may be biased *either* towards heads or tails. In many cases, a statistical test can be done either one-tailed or two-tailed; the two-tailed method often produces a *p*-value twice as big as the one-tailed method. Because small *p*-values are evidence against the null hypothesis, the one-tailed test seems to produce stronger evidence than its two-tailed counterpart. However, the advantage is largely illusory, as the example suggests.

Some experts have argued for one or the other type of test,[155] but a rigid rule is not required if significance levels are used as guidelines rather than as mechanical rules for statistical proof. One-tailed tests often make it easier to reach a threshold such as 5%, at least in terms of appearance. However, if we recognize that 5% is not a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the *p*-value are made explicit.

How Many Tests Have Been Done?

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield "significant" findings, even when there is no real effect. To illustrate the point, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce 10 heads when tossed 10 times is $(1/2)^{10} = 1/1024$. Observing 10 heads in the first 10 tosses therefore would be strong evidence that the coin is biased. Nonetheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. Ten heads in the first ten tosses means one thing; a run of ten heads somewhere along the way to a few thousand tosses means quite another. A test—looking for a run of ten heads—can be repeated too often.

Artifacts from multiple testing are commonplace. Because research that fails to uncover significance often is not published, reviews of the literature may

155. *See, e.g.*, Jones v. Novartis Pharm. Corp., 235 F. Supp. 3d 1244, 1288–89 (N.D. Ala. 2017); United States v. State of Del., 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 WL 609331, *10 n.27 (D. Del. 2004). According to formal statistical theory, the choice between one tail or two can sometimes be made by considering the exact form of the alternative hypothesis (see section titled "What Are the Rival Hypotheses?" below). *But see* Freedman et al., *supra* note 14, at 547–50. One-tailed tests at the 5% level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. One-tailed tests are also called one-sided (with no pejorative intent); two-tailed tests are two-sided.

produce a misleadingly large number of studies finding statistical significance.[156] The many other nonsignificant tests are not part of the literature; but they have been carried out, and like the omitted tails in the coin flipping example, they have implications for interpreting the studies that have been published. Thus, multiple testing is a factor in the failure of many scientific studies to replicate.[157]

Even a single researcher may examine so many different relationships that a few will achieve statistical significance by mere happenstance. For example, the researcher could choose a range of different outcome variables or different explanatory variables. Almost any large dataset—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by diligent search. Having detected the pattern, the analyst can perform a statistical test for it, blandly ignoring the search effort.[158] Statistical significance is bound to follow.[159]

There are statistical methods for dealing with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.[160] In addition, alternative approaches that focus on controlling the false discovery rate (the fraction of significant results expected to be false positives) are gaining in popularity. However, no general solution is known for handling multiple comparisons, and the existing methods would be of little help in the typical case where analysts have tested and rejected a variety of models before arriving at the one considered the most satisfactory (see section titled "Correlation and Regression" below on regression models). In these situations, researchers should be disclosing their multiple comparisons,[161]

156. *E.g.*, Philippa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 Lancet 867 (1991), https://doi.org/10.1016/0140-6736(91)90201-7; John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 JAMA 281 (1998); Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 New Eng. J. Med. 426 (1987), https://doi.org/10.1056/NEJM198708133170706.

157. *See* section titled "Tests or Interval Estimates?" above.

158. The practice has been denominated HARKing. Norbert L. Kerr, *HARKing: Hypothesizing After the Results Are Known*, 2 Personality & Social Psych. Rev. 196 (1998), https://doi.org/10.1207/s15327957pspr0203_4.

159. Searching for significance in this way has been called data mining, data dredging, data snooping, selection bias, and, more recently, *p*-hacking. Megan L. Head et al., *The Extent and Consequences of P-Hacking in Science*, 13 PLOS Biology e1002106 (2015), https://doi.org/10.1371/journal.pbio.1002106.

160. *See, e.g.,* Sandrine Dudoit & Mark J. van der Laan, Multiple Testing Procedures with Applications to Genomis (2008); Kaye, *supra* note 121, at 61–63; Martin Krzywinski & Naomi Altman, *Points of Significance: Comparing Samples—Part II*, 11 Nature Methods 355 (2014); Pak C. Sham & Shaun M. Purcell, *Statistical Power and Significance Testing in Large-scale Genetic Studies*, 15 Nature Reviews Genetics 335 (2014). In *Karlo v. Pittsburgh Glass Works*, 849 F.3d 61 (3d Cir. 2017), the court of appeals held that *Daubert* does not require an expert to use a particular (and conservative) adjustment method to find that a company's reduction in force had a statistically significant disparate impact in three out of five overlapping subgroups of older workers. *Id.* at 83. For discussion of the case, *see* David H. Kaye, *Multiple Hypothesis Testing in* Karlo v. Pittsburgh Glass Works, Forensic Sci., Stat. & L., July 5, 2017, https://perma.cc/JTD5-FULS.

161. ASA Comm. on Professional Ethics, *supra* note 11, at 3.

and courts should not be overly impressed with claims that estimates are significant. Instead, they should be asking how analysts developed their models. Intuition may suggest that the more variables included in the model, the better. However, this idea often turns out to be wrong. Complex models and more variables may increase the chance of a spurious result (one that reflects only accidental features of the data). Standard statistical tests offer little protection against this possibility when the analyst has tried a variety of models before settling on the final specification.

What Are the Rival Hypotheses?

The *p*-value of a statistical test is computed on the basis of a model for the data: the null hypothesis. Usually, the test is made in order to argue for the alternative hypothesis: another model. However, on closer examination, both models may be unreasonable. A small *p*-value means something is going on besides random error. The alternative hypothesis should be viewed as one possible explanation, out of many, for the data.

In *Mapes Casino, Inc. v. Maryland Casualty Co.*,[162] the court recognized the importance of explanations that the proponent of the statistical evidence had failed to consider. In this action to collect on an insurance policy, Mapes sought to quantify its loss from theft. It argued that employees were using an intermediary to cash in chips at other casinos. The casino established that over an 18-month period, the win percentage at its craps tables was 6%, compared to an expected value of 20%. The statistics proved that *something* was wrong at the craps tables—the discrepancy was too big to explain as the product of random chance. But the court was not convinced by plaintiff's alternative hypothesis. The court pointed to other possible explanations (Runyonesque activities such as skimming, scamming, and crossroading) that might have accounted for the discrepancy without implicating the suspect employees.[163] Rejection of the null hypothesis did not leave the proffered alternative hypothesis as the only viable explanation for the data.[164]

162. 290 F. Supp. 186 (D. Nev. 1968).

163. *Id.* at 193. Skimming consists of "taking off the top before counting the drop," scamming is "cheating by collusion between dealer and player," and crossroading involves "professional cheaters among the players." *Id.* In plainer language, the court seems to have ruled that the casino itself might be cheating, or there could have been cheaters other than the particular employees identified in the case. At the least, plaintiff's statistical evidence did not rule out such possibilities. *Compare* EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 312 & n.9, 313 (7th Cir. 1988) (EEOC's regression studies showing significant differences did not establish liability because surveys and testimony supported the rival hypothesis that women generally had less interest in commission sales positions), *with* EEOC v. Gen. Tel. Co., 885 F.2d 575 (9th Cir. 1989) (unsubstantiated rival hypothesis of "lack of interest" in "nontraditional" jobs insufficient to rebut prima facie case of gender discrimination); *cf.* section titled "Is the Study Designed to Investigate Causation?" above (problem of confounding).

164. *E.g.*, Coleman v. Quaker Oats Co., 232 F.3d 1271, 1283 (9th Cir. 2000) (A disparity with a *p*-value of "3 in 100 billion" did not demonstrate age discrimination because "Quaker never

Bayesian Statistical Methods and Posterior Probabilities

Standard errors, *p*-values, and significance tests are common techniques for assessing the potential impact of random errors or chance events. These procedures rely on sample data and their use is justified primarily in terms of the operating characteristics of statistical procedures.[165] They are often referred to as frequentist procedures because of this reliance on properties of the procedures that would be observed in repeated samples. As indicated in previous sections, frequentist procedures do not allow for assertions regarding the probability that a particular hypothesis is correct or that a particular confidence interval contains the true parameter value, given the data. For example, a statistician may postulate that a coin is fair: There is a 50–50 chance of landing heads, and successive tosses are independent. This is an empirical statement—potentially falsifiable—about the coin. It is easy to calculate the chance that a fair coin will turn up heads in every one of the next 10 tosses: The answer is $(1/2)^{10} = 1/1024$. Therefore, observing 10 heads in a row brings into serious doubt the initial hypothesis of fairness.

But what of the converse probability: If the coin does land heads 10 times, what is the chance that it is fair?[166] To compute such converse probabilities, it is necessary to postulate initial probabilities that the coin is fair, as well as probabilities of unfairness to various degrees. In the Bayesian approach, probabilities represent subjective degrees of belief about hypotheses or causes rather than objective facts about observations. The observer must quantify beliefs about the chance that the coin is unfair to various degrees—in advance of seeing the data. For example, let p be the unknown probability that the coin lands heads. What is the chance that p exceeds 0.1? 0.6? These subjective probabilities, like the probabilities governing the tosses of the coin, are set up to obey the axioms of probability

contends that the disparity occurred by chance; just that it did not occur for discriminatory reasons. When other pertinent variables were factored in, the statistical disparity diminished and finally disappeared.").

165. Operating characteristics include the expected value and standard error of estimators, probabilities of error for statistical tests, and the like.

166. We call this a converse probability because it is of the form $P(H_0|\text{data})$ rather than $P(\text{data}|H_0)$; an equivalent phrase, "inverse probability," also is used. Treating $P(\text{data}|H_0)$ as if it were the converse probability $P(H_0|\text{data})$ is the transposition fallacy. For example, most U.S. senators are men, but few men are senators. (As of 2024, 75 of the 100 senators are men.) Consequently, there is a high probability (0.75) that an individual who is a senator is a man, but the probability that an individual who is a man is a senator is practically zero. For examples of the transposition fallacy in court opinions, see cases cited *supra* notes 126 & 136. The frequentist *p*-value, $P(\text{data}|H_0)$, is generally not a good approximation to the Bayesian $P(H_0|\text{data})$; the latter includes considerations of power and base rates.

theory.[167] The probabilities for the various hypotheses about the coin, specified before data collection, are called prior probabilities. Prior probabilities can be updated, using Bayes' rule, given data on how the coin actually falls. (The Appendix explains the rule.) In short, a Bayesian analysis involves posterior probabilities for various hypotheses about the coin, given the data. These posterior probabilities quantify the statistician's confidence in the hypothesis that a coin is fair.[168] Although such posterior probabilities relate directly to hypotheses of legal interest, they are necessarily subjective, for they reflect not just the data but also the subjective prior probabilities—that is, degrees of belief about hypotheses formulated prior to obtaining data.[169]

With the exceptions of parentage testing and so-called probabilistic genotyping software for interpreting complex DNA mixtures,[170] such analyses have rarely been used in court, and the question of their forensic value was aired primarily in the academic literature. Some statisticians favor Bayesian methods, and such methods have become increasingly popular for settings in which a number of parameters are to be estimated simultaneously (for example, in assessing the effectiveness of standardized test coaching programs offered at a number of different schools).[171] Some legal commentators have proposed using these methods in various settings.[172]

167. Bayesian procedures are sometimes defended on the ground that the beliefs of any rational observer must conform to the Bayesian rules. However, the definition of "rational" is purely formal. *See* Peter C. Fishburn, *The Axioms of Subjective Probability*, 1 Stat. Sci. 335 (1986); David Freedman, *Some Issues in the Foundation of Statistics*, 1 Found. Sci. 19 (1995), *reprinted in* Topics in the Foundation of Statistics 19 (Bas C. van Fraasen ed., 1997); David Kaye, *The Laws of Probability and the Law of the Land*, 47 U. Chi. L. Rev. 34 (1979).

168. Here, confidence has the meaning ordinarily ascribed to it, rather than the technical interpretation applicable to a frequentist confidence interval. Consequently, it can be related to the burden of persuasion. *See* David H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

169. *But see infra* note 226.

170. *See* Kaye, *supra* note 55.

171. Donald B. Rubin, *Estimation in Parallel Randomized Experiments*, 6 J. Educ. Stat. 377 (1981), https://doi.org/10.3102/10769986006004377. Many practicing statisticians are pragmatists, using whatever procedure they think is appropriate for the occasion.

172. *See* Christopher S. Elmendorf & Douglas M. Spencer, *Administering Section 2* of *the Voting Rights Act After* Shelby County, 115 Colum. L. Rev. 2143, 2215 (2015); David H. Kaye, *Forensic Statistics in the Courtroom*, *in* Handbook of Forensic Statistics 225–48 (David Banks et al. eds., 2021); Kaye et al., *supra* note 8, §§ 12.8.5 & 14.3.2; David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N.C. L. Rev. 425 (2009). In addition, as indicated in the Appendix, Bayes' rule is crucial in solving certain problems involving conditional probabilities of related events. For example, if the proportion of women with breast cancer in a region is known, along with the probability that a mammogram of an affected woman will be positive for cancer and that the mammogram of an unaffected woman will be negative, then one can compute the numbers of false-positive and false-negative mammography results that would be expected to arise in a population-wide screening program. Using Bayes' rule to diagnose a specific patient, however, is more problematic, because the prior probability that the patient has breast cancer may

Correlation and Regression

Regression models are used by many social scientists to infer causation from association. Such models have been offered in court to prove disparate impact in discrimination cases, to estimate damages in antitrust actions, and for many other purposes. The sections titled "Scatter Diagrams," "Correlation Coefficients," and "Regression Lines" cover some preliminary material, showing how scatter diagrams, correlation coefficients, and regression lines can be used to summarize relationships between variables.[173] The section titled "Statistical Models" explains the ideas of regression modeling and some of the pitfalls, particularly those arising in the use of regression models to infer causation.

Scatter Diagrams

The relationship between two variables can be graphed in a scatter diagram (also called a scatterplot or scattergram). We begin with data on income and education for a sample of 238 men, ages 25 to 34, residing in Kansas.[174] Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 5, the horizontal axis shows education, and the vertical axis shows income. Person A completed 12 years of schooling (high school) and had an income of $50,000. Person B completed 16 years of schooling (college) and had an income of $100,000.

not equal the population proportion. Nevertheless, to overcome the tendency to focus on a test result without considering the "base rate" at which a condition occurs, a diagnostician can apply Bayes' rule to plausible base rates before making a diagnosis. Finally, Bayes' rule also is valuable as a device to explicate the meaning of concepts such as error rates, probative value, and transposition. *See, e.g.,* David H. Kaye, The Double Helix and the Law of Evidence (2010); Kaye et al., *Wigmore*, *supra* note 44, § 7.3.2; David H. Kaye, *Digging into the Foundations of Evidence Law*, 115 Mich. L. Rev. 915 (2017).

173. The focus is on simple linear regression. *See also* Rubinfeld & Card, *supra* note 25, for further discussion of these ideas with an emphasis on econometrics.

174. These data are from a public-use data file of the 2021 American Community Survey and were obtained from the data.census.gov website of the Bureau of the Census, U.S. Department of Commerce (precise query: https://data.census.gov/mdat/#/search?ds=ACSPUMS1Y2021&vv=PERNP,%2aAGEP&cv=SEX&rv=ucgid,SCHL&wt=PWGTP&g=0400000US20). Income and education are self-reported. Income is censored at $200,000. Only positive income values are included. Education variables are recoded to yield years of education. Data are a 20% sample from the resulting dataset to make the figures easier to read. Both variables in a scatter diagram have to be quantitative (with numerical values) rather than qualitative (nonnumerical).

Figure 5. Plotting points in a scatter diagram.

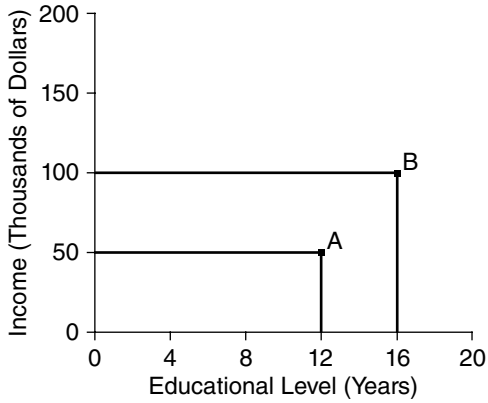
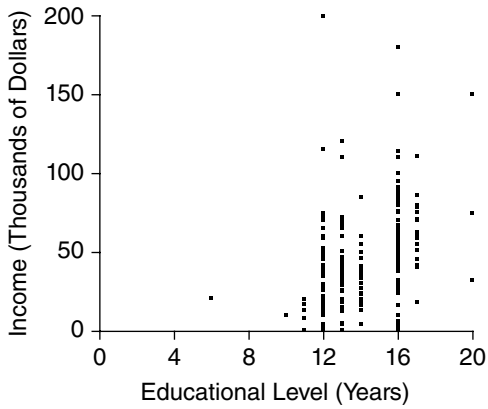


Figure 6 is the scatter diagram for the Kansas data. The diagram confirms an obvious point: There is a positive association between income and education. In general, persons with a higher educational level have higher incomes. However, there are many exceptions to this rule, and the association is not as strong as one might expect.

Figure 6. Scatter diagram for income and education: men ages 25 to 34 in Kansas.

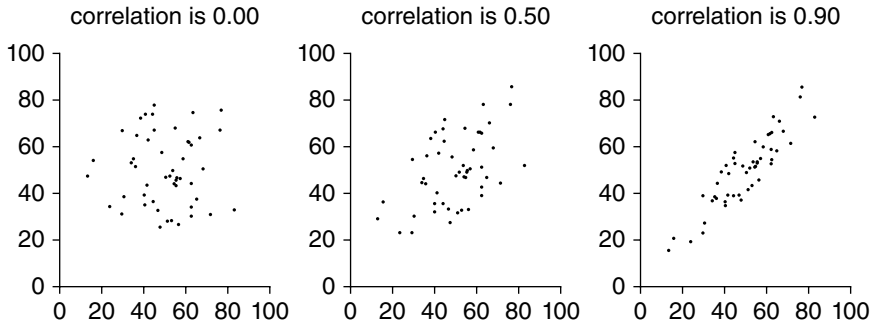


Correlation Coefficients

Two variables are positively correlated when their values tend to go up or down together, such as income and education in Figure 6. The correlation coefficient

(usually denoted by the letter *r*) is a single number that reflects the magnitude of a linear association and whether it is positive or negative. Figure 7 shows *r* for three scatter diagrams: In the first, there is no association; in the second, the association is positive and moderate; in the third, the association is positive and strong. Moving across these diagrams, the clouds of dots are increasingly clustered around a straight line that could be drawn through the cloud.

Figure 7. The correlation coefficient measures the sign of a linear association and its strength.



A correlation coefficient of 0 indicates no linear association between the variables. The maximum value for the coefficient is +1, indicating that all the dots fall exactly on a straight line that slopes up. Sometimes, there is a negative association between two variables: Large values of one tend to go with small values of the other. The age of a car and its fuel economy in miles per gallon illustrate the idea. Negative association is indicated by negative values for *r*. The extreme case is an *r* of –1, indicating that all the points in the scatter diagram lie on a straight line that slopes down.

Weak associations are the rule in the social sciences. In Figure 6, the correlation between income and education is about 0.4. The correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the correlation between LSAT scores and first-year grades is generally about 0.4.[175] The correlation between heights of fraternal twins is roughly 0.5. The correlation between heights of identical twins is about 0.9.[176]

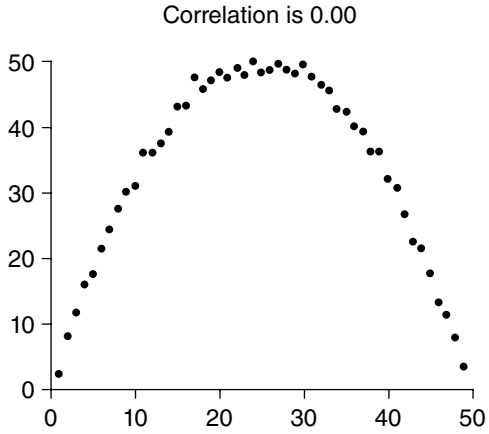
175. Law School Admissions Council, Summary of 2017, 2018, and 2019 LSAT Correlation Study Results, https://perma.cc/452D-JVNP. Adjusting for range restriction boosts the median correlations to 0.44 and 0.61. *Id.*

176. G. Frederiek Estourgie-van Burk et al., *Body Size in Five-year-old Twins: Heritability and Comparison to Singleton Standards*, 9 Twin Research & Hum. Genetics 646, 649 tbl. 2 (2006), https://doi.org/10.1375/183242706778553417 (reporting correlations of about 0.59 and 0.94 for the two types of

Is the Association Linear?

The correlation coefficient has a number of limitations, to be considered in turn. The correlation coefficient is designed to measure linear association. Figure 8 shows a strong nonlinear pattern with a correlation close to zero. The correlation coefficient is of limited use with a nonlinear relationship.

Figure 8. A strong nonlinear association with a correlation coefficient close to zero. The correlation coefficient only measures the degree of linear association.



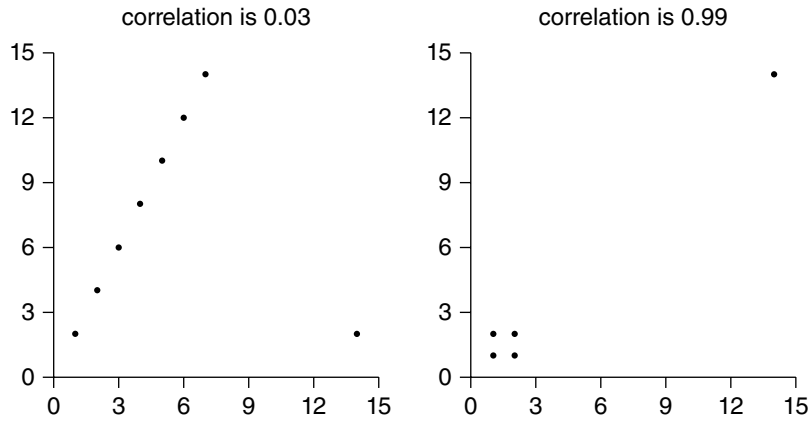
Do Outliers Influence the Correlation Coefficient?

The correlation coefficient can be distorted by outliers—a few points that are far removed from the bulk of the data. The left-hand panel in Figure 9 shows that one outlier (lower right-hand corner) can reduce a perfect correlation to nearly nothing. Conversely, the right-hand panel shows that one outlier (upper right-hand corner) can raise a correlation of zero to nearly one.[177] If there are extreme outliers in the data, the correlation coefficient is unlikely to be meaningful.

twins); Matthew C. Keller et al., *The Genetic Correlation Between Height and IQ: Shared Genes or Assortative Mating?*, 10 PLOS Genetics e1004329 tbl. 2 (2013), https://doi.org/10.1371/journal.pgen.1003451 (0.46 and 0.87); Jon Martin Sundet et al., *Resolving the Genetic and Environmental Sources of the Correlation Between Height and Intelligence: A Study of Nearly 2600 Norwegian Male Twin Pairs*, 8 Twin Research & Hum. Genetics 307, 309 tbl. 2 (2005), https://doi.org.10.1375/1832427054936745 (0.54 and 0.92).

177. *Cf.* David H. Kaye, *The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 Va. L. Rev 1933 (2001) (describing a simple linear regression in an antitrust case in which "[t]he difference between a finding by [plaintiff's expert] of several hundred

Figure 9. The correlation coefficient can be distorted by outliers.



Does a Confounding Variable Influence the Coefficient?

The correlation coefficient measures the association between two variables. Researchers—and the courts—are usually more interested in causation. Causation is not the same as association. The association between two variables may be driven by a lurking variable that has been omitted from the analysis (see section titled "Is the Study Designed to Investigate Causation?" above). For an easy example, there is an association between shoe size and vocabulary among schoolchildren. However, learning more words does not cause the feet to get bigger, and larger feet do not make children more articulate. In this case, the lurking variable is easy to spot—age. In more realistic examples, the lurking variable may be harder to identify.

In statistics, lurking variables are called confounders or confounding variables. Association may reflect causation, but a large correlation coefficient is not enough to warrant causal inference. A large value of r means only that the dependent variable marches in step with the independent one: Possible reasons include causation, confounding, and coincidence. Multiple regression is one method that attempts to deal with confounders (see "Statistical Models" below).[178]

million dollars of damages and a finding of no damages is the inclusion in his model of a single anomalous data point"—a result that "would not be acceptable in an undergraduate econometrics class, let alone professional work" but that went undetected at trial) (quoting Brief of Amicus Curiae Dr. Daniel L. McFadden in Support of Defendants-Appellants at 15, 19, Conwood Co. v. United States Tobacco Co., 290 F.3d 768 (2002)).

178. *See also* Rubinfeld & Card, *supra* note 25. The difference between experiments and observational studies is discussed in the section titled "Descriptive Surveys and Censuses" above.

Regression Lines

The regression line can be used to describe a linear trend in the data. The regression line for income on education in the Kansas sample is shown in Figure 10. The height of the line estimates the average income for a given educational level. For example, the average income for people with 12 years of education is estimated at $35,600, indicated by the height of the line at 12 years. The average income for people with 16 years of education is estimated at $58,400.

Figure 10. The regression line for income on education and its estimates.

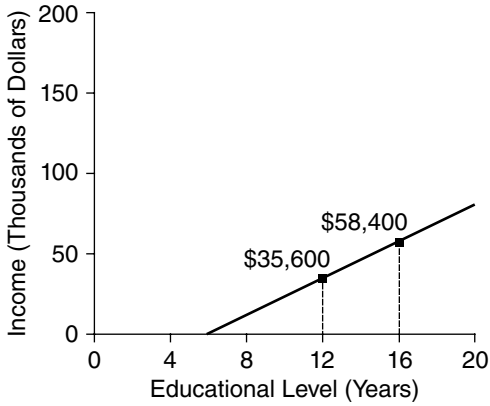
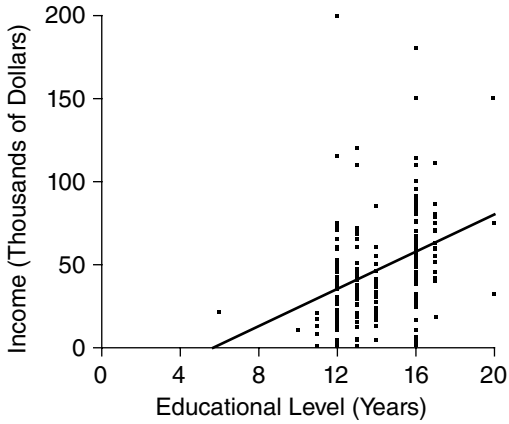


Figure 11 shows the points for income and education that were plotted in Figure 6 with the regression line superimposed. Many straight lines could have been drawn through the data points. Some would "fit" the data better than others. The regression line in Figure 11 comes from a method known as "ordinary least squares" (OLS). Every point in the scatterplot lies some distance directly above or below the line. We can square these distances and add them up. The OLS regression line is the one line for which this sum of the squared distances is the smallest. It shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated with a change in another variable (education).

What Are the Slope and Intercept?

The regression line can be described in terms of its intercept and slope. Often, the slope is the more interesting statistic. In Figure 10, the slope is $5,700 per year. On average, each additional year of education is associated with an additional $5,700 of income. Next, the intercept is –$32,800. This is an estimate of

Figure 11. Scatter diagram for income and education, with the regression line indicating the trend.



the average income for the (hypothetical) population of persons with zero years of education.[179] In this case, the estimate is nonsensical (below zero!) because zero years of education is very far from the bulk of the data where income appears to grow linearly with increasing years of education.

The slope of the regression line has the same limitations as the correlation coefficient: (1) The slope may be misleading if the relationship is strongly nonlinear; and (2) the slope may be affected by confounders and outliers. With respect to (1), the slope of $5,700 per year in Figure 10 presents each additional year of education as having the same value, but some years of schooling surely are worth more and others less. With respect to (2), the association between education and income is no doubt causal, but there are other factors to consider, including family background. Compared to individuals who did not graduate from high school, people with college degrees usually come from richer and better-educated families. Thus, college graduates have advantages besides education. As statisticians

179. The regression line, like any straight line, has an equation of the form $y = a + bx$. Here, a is the intercept (the value of y when $x = 0$), and b is the slope (the change in y per unit change in x). In Figure 11, the intercept of the regression line is –$32,800 and the slope is $5,700 per year. The line estimates an average income of $58,400 for people with 16 years of education. This may be computed from the intercept and slope as follows:

$$-\$32{,}800 + (\$5{,}700 \text{ per year}) \times 16 \text{ years} = -\$32{,}800 + \$91{,}200 = \$58{,}400.$$

The slope b is the same anywhere along the line. Mathematically, that is what distinguishes straight lines from other curves. If the association is negative, the slope will be negative too. The slope is like the grade of a road, and it is negative if the road goes downhill. The intercept is like the starting elevation of a road, and it is computed from the data so that the line goes through the center of the scatter diagram, rather than being generally too high or too low.

might say, the effects of family background are confounded with the effects of education. Statisticians often use the guarded phrases "on average" and "associated with" when talking about the slope of the regression line. This is because the slope has limited utility when it comes to making causal inferences.

What Is the Unit of Analysis?

If association between characteristics of individuals is of interest, these characteristics should be measured on individuals. Sometimes individual-level data do not exist, but rates or averages for groups are available. "Ecological" correlations are computed from such rates or averages. These correlations generally overstate the strength of an association. For example, average income and average education can be determined for men living in each state and in Washington, D.C. The correlation coefficient for these 51 pairs of averages turns out to be 0.7. However, states do not go to school and do not earn incomes. People do. The correlation for income and education for men in the United States is only 0.4. The correlation for state averages overstates the correlation for individuals—a common tendency for ecological correlations.[180]

Ecological analysis is often seen in cases claiming dilution in voting strength of minorities. In this type of voting rights case, plaintiffs must prove three things: (1) the minority group constitutes a majority in at least one district of a proposed plan; (2) the minority group is politically cohesive—that is, votes fairly solidly for its preferred candidate; and (3) the majority group votes sufficiently as a bloc to defeat the minority-preferred candidate.[181] The first requirement is compactness; the second and third define polarized voting.

180. Correlations are computed from the March 2005 Current Population Survey for men ages 25–64. Freedman et al., *supra* note 14, at 149. The ecological correlation uses only the average figures, but within each state there is a lot of spread about the average. The ecological correlation smoothes away this individual variation. *Cf.* Gold et al., *supra* note 15 (suggesting that ecological studies of exposure and disease are "far from conclusive" because of the lack of data on confounding variables (a much more general problem) as well as the possible aggregation bias described here); David A. Freedman, *Ecological Inference and the Ecological Fallacy, in* 6 Int'l Encyclopedia of the Social and Behavioral Sciences 4027 (Neil J. Smelser & Paul B. Baltes eds., 2001).

181. *See* Thornburg v. Gingles, 478 U.S. 30, 50–51 (1986) ("First, the minority group must be able to demonstrate that it is sufficiently large and geographically compact to constitute a majority in a single-member district. . . . Second, the minority group must be able to show that it is politically cohesive. . . . Third, the minority must be able to demonstrate that the white majority votes sufficiently as a bloc to enable it . . . usually to defeat the minority's preferred candidate."). In subsequent cases, the Court has emphasized that these factors are not sufficient to make out a violation of section 2 of the Voting Rights Act. *E.g.*, Johnson v. De Grandy, 512 U.S. 997, 1011 (1994) ("*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court's examination of relevant circumstances was complete once the three factors were found to exist, or in the sense that the three in combination necessarily and in all circumstances

The secrecy of the ballot box means that polarized voting cannot be directly observed. Instead, plaintiffs in voting rights cases rely on ecological regression, with scatter diagrams, correlations, and regression lines to estimate voting behavior by groups and demonstrate polarization.[182] The unit of analysis typically is the precinct. For each precinct, public records can be used to determine the percentage of registrants in each demographic group of interest, as well as the percentage of the total vote for each candidate—by voters from all demographic groups combined. Plaintiffs' burden is to determine the vote by each demographic group separately.

Figure 12 shows how the argument unfolds. Each point in the scatter diagram represents data for one precinct in the 1982 Democratic primary election for auditor in Lee County, South Carolina. The horizontal axis shows the percentage of registrants who are white. The vertical axis shows the turnout rate for the white candidate. The regression line is plotted too. The slope would be interpreted as the expected increase in the proportion supporting the white candidate for each 1% increase in the proportion of white registrants in the precinct. The intercept then would be the expected support for the white candidate in a precinct with no white registrants (an all-Black precinct).[183] The validity of such estimates is contested in the statistical and legal literature.[184]

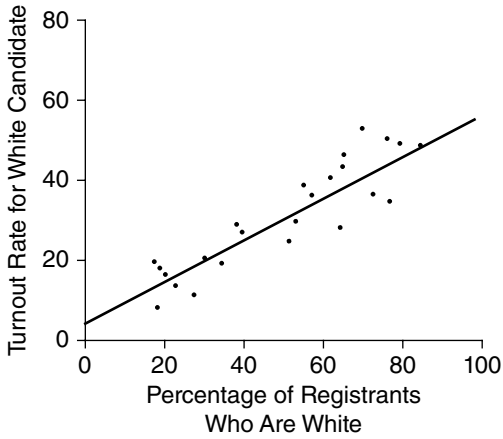
demonstrated dilution."). On the legal framework and its problems, *see* Christopher S. Elmendorf et al., *Racially Polarized Voting*, 83 U. Chi. L. Rev. 587 (2016); D. James Greiner, *Re-Solidifying Racial Bloc Voting: Empirics and Legal Doctrine in the Melting Pot*, 86 Ind. L.J. 447 (2011); Nicholas O. Stephanopoulos, *The Relegation of Polarization*, 83 U. Chi. L. Rev. Online 160, 168 (2017).

182. Less readily visualized procedures also are used. *See, e.g.,* Ecological Inference: New Methodological Strategies (Gary King et al. eds., 2004); Loren Collingwood et al., *The R Journal: eiCompare: Comparing Ecological Inference Estimates Across EI and EI:R×C*, R J., Dec. 2016, at 92, https://perma.cc/U5JH-GBEM. We confine our explanation to the simplest (bivariate linear) form of ecological regression. All the methods confront the same fundamental problem of inferring associations at the level of individuals from differences among groups that include heterogeneous subgroups of individuals. Given this common issue and objective, one might think that the phrase "ecological inference" would be a broad umbrella term, but especially in voting rights litigation, the phrase usually refers to one particular procedure that is more complex than traditional bivariate ecological regression. *See, e.g.,* Luna v. County of Kern, 291 F. Supp. 3d 1088, 1118 (E.D. Cal. 2018) (describing differences between ecological regression (ER) and "[e]cological inference ('EI'), developed by political scientist Gary King in 1997 [that] is 'similar to, but largely regarded as an improvement upon' the ER methodology endorsed in *Gingles*"). *But see* Elmendorf & Spencer, *supra* note 172, at 2180 (referring to "statistically tenuous techniques of ecological inference"); Greiner, *supra* note 181, at 449 (observing that "analyzing census data (usually) and precinct-level vote returns via a set of methods collectively called 'ecological inference' has been a staple since the mid 1980s").

183. By definition, the turnout rate equals the number of votes for the candidate, divided by the number of registrants; the rate is computed separately for each precinct. The intercept of the line in Figure 11 is 4%, and the slope is 0.52. Plaintiffs would conclude that only 4% of the voters in an all-Black precinct would be expected to vote for the white candidate, while 4% + 52% = 56% of the voters in an all-white precinct would be expected to vote for the white candidate, which demonstrates polarization.

184. *E.g.*, Moon Duchin & Douglas M. Spencer, *Models, Race, and the Law*, 130 Yale L.J. Forum 744 (2021); Elmendorf & Spencer, *supra* note 172, at 2203 ("Ecological inference as

Figure 12. Turnout rate for the white candidate plotted against the percentage of registrants who are white. Precinct-level data, 1982 Democratic primary for auditor, Lee County, South Carolina.[185]



Statistical Models

Statistical models are widely used in the social sciences and in litigation. For example, the census suffers an undercount, more severe in certain places than others.

traditionally practiced relies on heroic assumptions, elides questions about statistical precision, and uses post-hoc corrections to paper over mathematically impossible results. . . .") (note omitted). For references to some of the statistical literature, see Collingwood et al., *supra* note 182, at 92–94. The use of ecological regression and other quantitative methods increased considerably after the Supreme Court noted in *Thornburg v. Gingles*, 478 U.S. 30, 53 n.20 (1986), that "[t]he District Court found both methods [extreme case analysis and bivariate ecological regression analysis] standard in the literature for the analysis of racially polarized voting." *See* Bruce M. Clarke & Robert Timothy Reagan, Federal Judicial Center, Redistricting Litigation: An Overview of Legal, Statistical, and Case-Management Issues (2002); D. James Greiner, *The Quantitative Empirics of Redistricting Litigation: Knowledge, Threats to Knowledge, and the Need for Less Districting*, 29 Yale L. & Policy Rev. 527 (2011); D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want to Be?*, 47 Jurimetrics J. 115, 117, 121 (2007).

Redistricting plans based predominantly on racial considerations are unconstitutional unless narrowly tailored to meet a compelling state interest. Shaw v. Reno, 509 U.S. 630 (1993). Whether compliance with the Voting Rights Act can be considered a compelling interest is an open question, but efforts to sustain racially motivated redistricting on this basis have not fared well before the Supreme Court. *See* Abrams v. Johnson, 521 U.S. 74 (1997); Shaw v. Hunt, 517 U.S. 899 (1996); Bush v. Vera, 517 U.S. 952 (1996); Abbott v. Perez, 585 U.S. 579 (2018); *but see* Alabama Legislative Black Caucus v. Alabama, 575 U.S. 254 (2015); Bethune-Hill v. Virginia State Bd. of Elections, 580 U.S. 178 (2017); Cooper v. Harris, 581 U.S. 285 (2017).

185. Data from James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589, 591 tbl.1 (1989).

If some statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in tax funds.[186] Other models purport to lift the veil of secrecy from the ballot box, enabling the experts to determine how minority groups have voted—a crucial step in voting rights litigation (see section titled "Regression Lines" above). This section discusses the statistical logic of regression models.

A regression model attempts to combine the values of certain variables (the independent variables) to get expected values for another variable (the dependent variable). The model can be expressed in the form of a regression equation. A simple regression equation has only one independent variable; a multiple regression equation has several independent variables. Coefficients in the equation will be interpreted as showing the effects of changing the corresponding variables. This is justified in some situations, as the next example demonstrates.

Hooke's law (named after Robert Hooke, England, 1653–1703) describes how a spring stretches in response to a load: Strain is proportional to stress. To verify Hooke's law experimentally, a physicist will make a number of observations on a spring. For each observation, the physicist hangs a weight on the spring and measures its length. A statistician could develop a regression model for these data:

$$length = a + (b \times weight) + \varepsilon \tag{1}$$

The error term, denoted by the Greek letter ε (epsilon) is needed because measured length will not be exactly equal to $a + (b \times \text{weight})$. If nothing else, error in measuring the spring's length must be reckoned with.[187] The model takes ε as "random error"—behaving like draws made at random with replacement from a box of tickets. Each ticket shows a potential error, which will be realized if that ticket is drawn. The average of the potential errors in the box is assumed to be zero.[188]

Equation (1) has two parameters, a and b. These constants of nature characterize the behavior of the spring: a is length under no load, and b is elasticity (the increase in length per unit increase in weight). By way of numerical illustration, suppose a is 400 and b is 0.05. If the weight is 1, the length of the spring is expected to be

$$400 + (0.05 \times 1) = 400.05.$$

186. *See* Brown et al., *supra* note 36.

187. We will assume that the physicist uses a perfectly accurate and unchanging set of standard weights.

188. In standard statistical terminology, the process of measuring length here is unbiased.

If the weight is 3, the expected length is

$$400 + (0.05 \times 3) = 400 + 0.15 = 400.15.$$

In either case, the actual length will differ from expected, by a random error ε.

In the simplest situation, the ε's for different observations on the spring are assumed to be independent and identically distributed, with a mean of zero. "Independent" means that, for any two observations using the same weight, the chances for one ε do not depend on outcomes for the other. If the errors are like draws made at random with replacement from a box of tickets, as we assumed earlier, they are independent—the box will not change from one draw to the next. "Identically distributed" means that the chance behavior of the two ε's is the same: They are drawn at random from the same box.

The parameters a and b in equation (1) are not directly observable, but they can be estimated by the method of least squares.[189] Statisticians often denote estimates by hats. Thus, $\hat{a}$ is the estimate for a, and $\hat{b}$ is the estimate for b. The values of $\hat{a}$ and $\hat{b}$ are chosen to minimize the sum of the squared prediction errors. These errors are also called residuals. They measure the difference between the actual length of the spring and the predicted length, the latter being $\hat{a} + \hat{b} \times$ weight:

$$\textit{actual length} = \hat{a} + (\hat{b} \times \textit{weight}) + \textit{residual} \qquad (2)$$

Of course, no one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box.[190] In short, the statistical model corresponds rather closely to the empirical phenomenon.

Equation (1) is a statistical model for the data, with unknown parameters a and b. The error term ε is not observable. The model is a theory—and a good one—about how the data are generated. By contrast, equation (2) is a regression equation that is fitted to the data: The intercept $\hat{a}$, the slope $\hat{b}$, and the residual can all be computed from the data. The results are useful because $\hat{a}$ is a good estimate for a, and $\hat{b}$ is a good estimate for b. (Similarly, the residual is a good approximation

189. *See* section titled "Regression Lines" above. It might seem that a is observable; after all, we can measure the length of the spring with no load. However, the measurement is subject to error, so we observe not a, but $a + \varepsilon$. *See* equation (1). The parameters a and b can be estimated, even estimated very well, but they cannot be observed directly. The least squares estimates of a and b are the intercept and slope of the regression line. *See* section titled "What Are the Slope and Intercept?" above *and* Freedman et al., *supra* note 14, at 208–10. The method of least squares was developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits.

190. This is the Gauss model for measurement error. *See* Freedman et al., *supra* note 14, at 450–52.

to ε.) Without the theory, these estimates would be less useful. Is there a theoretical model behind the data processing? Is the model justifiable? These questions can be critical when it comes to making statistical inferences from the data.

These points apply to more complicated statistical models. The simple linear regression model exemplified in equation (1) can be extended in many ways. If another variable might contribute to or be associated with a response, an additional term (*constant* × *variable*) can be added to the right-hand side of the equation. Furthermore, the relationship between the independent and dependent variables need not be linear. For instance, to model measurements of the distance d a ball dropped from the Leaning Tower of Pisa falls in a few seconds, Newton's laws imply that we should use a model with the time variable squared.[191] Also, variables that are dichotomous rather than continuous (such as whether a mouse will develop cancer after exposure to a food additive) can be part of a regression model.[192]

In social science and legal applications, such advanced statistical models often are invoked—even when they lack an independent theoretical basis. Hooke's law—incorporated in equation (1)—is relatively easy to test experimentally. For something like the salaries that an employer pays to workers, validation of a proposed relationship between the dependent variable and the independent ones would be difficult to validate experimentally. When expert testimony relies on statistical models, the court may well inquire, what are the assumptions behind the model, and why do they apply to the case at hand? The assumptions being questioned can include the choice of variables in the regression and the nature of the assumed relationship. It is important to distinguish between two situations:

- The nature of the relationship between the variables is known and regression is being used to make quantitative estimates of parameters in that relationship, and
- The nature of the relationship is largely unknown and regression is being used to determine the nature of the relationship—or indeed whether any relationship exists at all.

Regression was developed to handle situations of the first type, with Hooke's law being an example. The basis for the second type of application is analogical, and the tightness of the analogy is an issue worth exploration.[193]

191. The model would be $distance = a + (b \times seconds^2) + \varepsilon$ instead of $distance = a + (b \times seconds) + \varepsilon$.

192. Such extensions and modifications of simple linear regression are introduced in the *Reference Guide on Multiple Regression and Advanced Statistical Models*; see also Andrew Gelman et al., *Regression and Other Stories* (2020).

193. *See, e.g.,* David A. Freedman, Statistical Models and Causal Inference: A Dialogue with the Social Sciences (David Collier et al. eds., 2009); Andrew Gelman, *Review Essay: Causality and Statistical Learning*, 117 Am. J. Sociology 955 (2011).

For most questions of legal interest, a wide variety of models can be used. This is only to be expected, because the science does not dictate specific equations and causal relationships. In a strongly contested case, each side will have its own model (series of models), presented by its own expert. The experts often reach opposite conclusions. The dialogue might continue with an exchange about which models produce admissible results, or more convincing ones. Although assumptions about the form of the model or the error component are challenged in court from time to time, arguments more commonly revolve around the choice of variables. One model may be questioned because it omits variables that arguably should be included—for example, skill levels or prior evaluations in an employment discrimination case.[194] Another model may be challenged because it includes tainted variables reflecting past discriminatory behavior by the firm.[195] One expert may emphasize how remarkably well the fitted regression curve or surface fits; another expert may reply that the model, especially a complicated one, is ad hoc—it is tailored to the data in question and cannot be relied on to work well with other data arising from the same data-generating process.[196] The court or jury must decide which model—if any—fits the occasion.[197]

The frequency with which regression models are used is no guarantee that they are the best choice for any particular problem.[198] Indeed, from one perspective, a regression or other statistical model may seem to be a marvel of mathematical rigor. From another perspective, the model could be a set of assumptions,

194. *E.g.*, Bazemore v. Friday, 478 U.S. 385 (1986); *In re* Linerboard Antitrust Litig., 497 F. Supp. 2d 666 (E.D. Pa. 2007).

195. *E.g.*, McLaurin v. Nat'l R.R. Passenger Corp., 311 F. Supp. 2d 61, 65–66 (D.D.C. 2004) (holding that the inclusion of two allegedly tainted variables was reasonable in light of an earlier consent decree); *see also In re* Urethane Antitrust Litig., 166 F. Supp. 3d 501, 506 (D.N.J. 2016) ("courts have declined to fault a regression model for excluding variables that could have been manipulated by a defendant in furtherance of an antitrust conspiracy").

196. *E.g.*, Students for Fair Admissions, Inc. v. Univ. of N.C., 567 F. Supp. 3d 580, 622 (M.D.N.C. 2021) ("both experts agree that the degree of overfit is a factor they must take into account, the parties disagree on how to measure it"), *rev'd*, 600 U.S. 181 (2023); *In re* Urethane Antitrust Litig., 166 F. Supp. 3d at 505 (predictions were admissible despite the argument "that the models in this case were able to accurately match prices during the benchmark period not because they are reliable, but because they are 'overfit'").

197. *E.g.*, Chang v. Univ. of R.I., 606 F. Supp. 1161, 1207 (D.R.I. 1985) ("it is plain to the court that [defendant's] model comprises a better, more useful, more reliable tool than [plaintiff's] counterpart"); Presseisen v. Swarthmore Coll., 442 F. Supp. 593, 619 (E.D. Pa. 1977) ("[E]ach side has done a superior job in challenging the other's regression analysis, but only a mediocre job in supporting their own . . . and the Court is . . . left with nothing."), *aff'd*, 582 F.2d 1275 (3d Cir. 1978).

198. *See, e.g., In re* Urethane Antitrust Litig., 166 F. Supp. 3d at 504–05 (although "there are an abundance of judicial decisions supporting the premise that regression models can be a reliable tool for measuring damages in an antitrust case," plaintiff's "regression models must stand on their own two feet").

supported only by the say-so of the testifying expert. Intermediate judgments are also possible.[199]

Data Science and Statistical Machine Learning

"Statistical science" has been described as "the discipline of learning about the world from data,"[200] and "statistics" as the art of "learning from data."[201] Yet, there are now journals,[202] university degree programs,[203] and job positions[204] dedicated—not to statistics—but to "data science." Spurred by the availability of large and heterogeneous data collections, computer-intensive algorithmic procedures for making predictions and classifications are increasingly applied in science, business, and law enforcement.[205] These procedures have often been termed "analytics," "artificial intelligence," and "machine learning." How might testimony and reports from "data scientists" differ from more traditional data analyses of statisticians? What statistical considerations or methods should be used to validate the output from machine learning approaches? To supply background information for approaching these questions, this section describes the

199. *See, e.g.*, David W. Peterson, *Reference Guide on Multiple Regression*, 36 Jurimetrics J. 213, 214–15 (1996) (review essay); *see supra* note 25 for references to a range of academic opinion. More recently, some investigators have turned to graphical models. However, these models have serious weaknesses of their own. *See, e.g.,* David A. Freedman, *Graphical Models for Causation, and the Identification Problem*, 28 Evaluation Rev. 267 (2004), https://doi.org/10.1177/0193841X04266432.

200. Spiegelhalter, *supra* note 95, at 404 (Basic Books edition).

201. This is the subtitle of the Penguin Books edition of Speigelhalter, *supra* note 95.

202. The *Harvard Data Science Review* (https://perma.cc/T87V-VEP3), for example, "aim[s] to publish content that help define and shape data science as a scientifically rigorous and globally impactful multidisciplinary field based on the principled and purposed production, processing, parsing, and analysis of data."

203. U.S. News & World Rep., Best Undergraduate Data Science Programs, https://www.usnews.com/best-colleges/rankings/computer-science/data-analytics-science.

204. The Bureau of Labor Statistics has an occupational category for "data scientists" that excludes statisticians. U.S. Bureau of Labor Statistics, Occupational Employment and Wages, May 2021, https://www.bls.gov/oes/current/oes152051.htm.

205. Applications of particular legal interest include predicting criminal behavior at the individual level (*see, e.g.,* Richard A. Berk & J. Bleich, *Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment*, 12 Criminology & Public Policy 513 (2013), https://doi.org/10.1111/1745-9133.12407, classifying trace evidence such as toolmarks according to their source (*see, e.g.*, ANSI/ASB 062, Standard for Topography Comparison Software for Toolmark Analysis (2021)), and forensic speaker recognition (*see, e.g.,* Geoffrey Stewart Morrison & William C. Thompson, *Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony*, 18 Colum. Sci. & Tech. L. Rev. 326, 345–46 (2017); Zhongxin Bai & Xiao-Lei Zhang, *Speaker Recognition Based on Deep Learning: An Overview*, 140 Neural Networks 65 (2021), https://doi.org/10.1016/j.neunet.2021.03.004.

general nature of "data science" and sketches the broad outlines of the statistical thinking that is crucial to evaluating results from machine learning.[206]

What Is Data Science?

"There is a wide variety of definitions and criteria for what constitutes data science," making the term something of "a buzzword."[207] Typical definitions, such as "a collection of techniques used to extract value from data [that] rely on finding useful patterns, connections, and relationships within data,"[208] do not distinguish data science from applied statistics.[209] For centuries, statisticians have been concerned with discerning patterns and regularities in natural and social processes and with predicting uncertain outcomes.

Nonetheless, data science has emerged as *a combination* of statistics, computing, and informatics ideas about the collection, storage, analysis, visualization, and reporting of data. Advances in technology, computing, and data-storage capacity have produced a data explosion in commerce (partly driven by online shopping), science (massive sky surveys, environmental and climate data, genomic information, and much more), and digitalized voice, text, and pictures collected to support speech and image recognition.[210] The "data science" that "extracts value" from these new troves of data is a conglomeration of mathematically grounded modeling methods from traditional statistics and highly pragmatic procedures for making predictions and classifications. The latter methods often fall under the rubric of "machine learning" and are sketched in the next section.

What Is Machine Learning?

"Machine learning" refers to methods for using data to classify objects or patterns into categories, and to produce predictions of future outcomes or events. The

206. For a general discussion of AI and related societal implications, see James E. Baker & Laurie N. Hobert, *Reference Guide on Artificial Intelligence*, in this manual.

207. Vijay Kotu & Bala Deshpande, Data Science: Concepts and Practice 1 (2d ed. 2019).

208. *Id.*; *see also* Bradley Efron & Trevor Hastie, Computer Age Statistical Inference: Algorithms, Evidence, and Data Science 451 (2016) (noting that "[t]he data science association defines a practitioner as one who 'uses scientific methods to liberate and create meaning from raw data'").

209. *See* Matthew A. Jay & Mario Cortina Borja, *Quomoto Dicitur "Data Science" Latine*, Significance, Aug. 2020, at 26.

210. Initially the increased amount of data was heralded as the dawn of a Big Data era involving Big Data specialists. *See* Francis X. Diebold, *What's the Big Idea? "Big Data" and Its Origins*, Significance, Feb. 2021, at 36–37. "Big" can refer to the number of units (people, social media posts, galaxies, etc.) in a dataset ("examples" in a database). It also can refer to the number of variables for each unit (the "features" of the examples). Genomes are big—they have billions of base pairs—even if they come from only a handful of people.

equations or procedures rely on computers. Of course, classification and prediction are tasks that humans do, either intuitively or with experience and expert training. Thus, machine learning (ML) is a subfield of "artificial intelligence" (AI), which seeks to build machines that perform tasks we associate with intelligent agents.[211]

ML is closely related to statistics. Some methods that are packaged as ML or AI are nothing fundamentally new.[212] For example, regression models can play a role. An easily understood example—predicting the first-year grade-point average (GPA) of a student entering law school from only one variable—illustrates some of the terminology. The school knows the LSAT score and the subsequent first-year GPA of every student in last year's class. It can make predictions by fitting a straight line to those data points. As explained in the sections titled "Correlation and Regression" and "What Inferences Can Be Drawn from the Data?" above, the fitted line could have the mathematical form $GPA = \hat{a} + \hat{b} \times LSAT$, where $\hat{a}$ and $\hat{b}$ are the line's intercept and slope as estimated from the data.[213] The estimates could come from equations for the values that minimize the sum of the squared differences between each data point and the fitted line. (That criterion, in turn, can be justified by a regression model in which $GPA = a + b \times LSAT + \varepsilon$, where a and b are the unknown parameters and ε is a normally distributed error term.) With the fitted equation in place (and under the assumption that the relationship between the variables for last year's students will continue to apply), the prediction of an applicant's

211. AI is extremely broad in its scope and tools, with no single, commonly accepted definition. *See, e.g.,* Stuart J. Russell & Peter Norvig, Artificial Intelligence: A Modern Approach (3d ed. 2010). Indeed, "so far at least, all the candidates [for a definition] have substantial flaws." Richard A. Berk, *Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement*, 4 Ann. Rev. Criminology 209, 211 (2021). For a time, AI focused on building symbolic representations of the world and developing rule- or case-based systems that could reason based on those representations. *See, e.g.,* Philip Leith, *The Rise and Fall of the Legal Expert System*, 1 European J. L. & Tech. No. 1 (2010), https://perma.cc/E8MF-JTFU. Although statisticians and machine-learning researchers do not think of their work as necessarily being motivated by traditional AI goals such as elucidating how the brain processes information, the statistical and algorithmic methods have produced automated systems that work well for narrow tasks. These accomplishments are now designated "narrow" or "weak" AI. However, "the artificial intelligence of today is computer code written to accomplish a specific empirical task. Restated, it is just a computational procedure. In law enforcement settings, these procedures are by and large enhancements of activities that humans are already performing. Arguably, computers can do them faster and more accurately, but the internal operations are far removed from the way humans actually think. It is not even clear that the term intelligence applies." Berk, *supra*, at 212.

212. Susan Athey & Guido W. Imbens, *Machine Learning Methods that Economists Should Know About*, 11 Ann. Rev. Econ. 685, 689 (2019), https://doi.org/10.1146/annrev-economics-080217-053433 (observing that "[o]ne source of confusion is the use of new terminology in ML for concepts that have well-established labels in the older literatures" and providing examples in the context of regression analysis); Berk, *supra* note 211, at 223–24 (some crime-forecasting procedures presented as ML "are a rebranding of statistical tools developed more than 50 years ago").

213. Additional predictors and other functional forms could be explored with the statistical procedures mentioned in the *Reference Guide on Multiple Regression and Advanced Statistical Models.*

score is simple arithmetic—one multiplication and one addition. The calculations, including the estimates of *a* and *b*, could be done by hand for a single law school, or the LSAT–GPA data could be entered into a spreadsheet or a statistical package of software that would generate the prediction equation and any desired prediction based on it. Next year, the machine will have new "training" data (another set of actual GPAs and LSATs) from which it can "learn" to make predictions (from a regression equation with updated, estimated values for the parameters *a* and *b*).

Prediction by simple linear regression is so well established it is unlikely to be packaged as machine learning. That phrase is more likely to be encountered with procedures that depart from such explicit modeling and that are said to be "algorithmic."[214] The word itself is not very revealing. Algorithms are step-by-step instructions for computing something. They can be written down in English or another natural language, or depicted in a flow chart. Then they can be implemented on a computer after they are converted into a specific computer programming language ("code").

The regression model-based procedure for predicting GPA entails two parts that can be called algorithmic. First, there was an algorithm for finding the particular regression line.[215] Second, this line enables us to write down an algorithm for calculating the predicted value. This second algorithm might go as follows: (1) input the applicant's LSAT score; (2) multiply it by the value $\hat{b}$ found from all the data; add the value found for $\hat{a}$ to the result; (3) output the result of step (2) for the predicted GPA.

Because algorithms underlie all computation, it is not the existence of algorithms that makes ML distinctive. It is the willingness to develop algorithms for predictions or classifications without necessarily positing a statistical model of how the data arise. Statisticians historically relied on statistical models with a relatively small number of interpretable parameters (like the intercept *a* and the slope *b* of the simple regression line) and used manageably small datasets to estimate the values of these parameters. But models with many parameters are characteristic of ML.[216]

To convey the rough idea, let's imagine that instead of imposing the linear regression model on the LSAT–GPA data, we used the following algorithm: (1) compute the average GPA for the students with LSATs ranging from 120–130,

214. *E.g.*, Berk, *supra* note 211, at 12 ("An essential feature of artificial narrow intelligence is that it depends on algorithms, not models."); Moritz Hardt & Benjamin Recht, Patterns, Predictions, and Actions: Foundations of Machine Learning 11 (2020) ("Machine learning is to a large extent the study of algorithmic prediction."); Efron & Hastie, *supra* note 208, at 451 ("the emphasis is on the algorithmic processing of large data sets for the extraction of useful information, with the prediction algorithms as exemplars").

215. The equations for estimating *a* and *b* in the regression model are solved by a particular series of additions, subtractions, and multiplications that are guaranteed to give the specific values $\hat{a}$ and $\hat{b}$ that minimize the sum of the squared deviations between each data point and the overall fitted line.

216. For discussion of how such procedures relate to more traditional statistical methods, *see Special Issue: Commentaries on Breimen's Two Cultures Paper*, 7 Observational Stud. 1–234 (2021).

130–140, and so on through 170–180; (2) plot these averages as heights at the midpoints of the intervals (125, 135, . . . , 175); (3) connect them to form a jagged line, and (4) use this jagged line to make predictions. Of course, one might ask why widths of 10 LSAT points should be used. Why not use shorter or longer segments for computing the means? An optimization procedure that reserves some of the data to try out different possibilities could help identify which width to use.[217] The approach resembles engineering by trial and error.[218]

Our jagged line would not be used in practice. There are better procedures for fitting a curve to a cloud of data points. The slice-compute-and-connect algorithm is here only to illustrate the *idea* of a highly data-driven procedure that leads to an algorithm (corresponding to the resulting jagged line) for making the predictions. The jagged line would hug the data more tightly than the single regression line with only two parameters. Maybe it would produce better predictions.

A number of important ML procedures do not yield any explicit equation (like the single regression line or the more complex jagged line) for making predictions or classifications. The two stages are done together, and the relevant variables are not necessarily pre-defined by the developer of the prediction model. Also, the features of the data that have the most influence on the output often are unknown.

What Statistical Questions Arise with Machine Learning Studies?

Although machine-learning advocates sometimes advertise that their methods make very few assumptions and provide great flexibility, they are still subject to the statistical and logical issues discussed in previous sections. Data quality,

217. We could reserve a random 10% of the data and develop the jagged line on the remaining 90% using a width of, say, two LSAT points. Then we check how well the fitted line works on the reserved 10% by finding the correlation between the fitted line's predictions and the reserved data points. We do this repeatedly, say 10 times, to get an average correlation for the jagged line with the two-point width. Then we return to the full training set and repeat this 10-fold train-test process for other values of the width. Finally, we pick the width with the highest correlation ("cross-validity") and apply it to the full training set (100% of the data). That gives the prediction line.

218. Moritz Hardt & Benjamin Recht, Patterns, Predictions, and Actions: Foundations of Machine Learning 146 (2020) ("Methodologically, much of modern machine learning practice rests on a variant of *trial and error*, which we call the *train-test paradigm*. Practitioners repeatedly build models using any number of heuristics and test their performance to see what works. Anything goes as far as training is concerned, subject only to computational constraints, so long as the performance looks good in testing.").

robustness, and transparency are a few of the issues that can arise.[219] We consider these three in turn.[220]

Is the Dataset Appropriate and of Sufficient Quality?

To begin with, as with traditional statistical methods, care is needed in the assembly of the data used to build and test the algorithms. Algorithms must be developed on data that are accurate and representative of the type of data to which they will be applied. For example, a study on diagnosing autism using ML excluded the data corresponding to borderline cases of autism. That made the test set unrepresentative of the general population; moreover, when a dataset that included the difficult-to-diagnose cases was used, the accuracy declined.[221]

Is the Predictor or Classifier Robust?

Second, robustness is also a vital issue. The goal of ML is not merely to find some complex pattern in the data; it is to discover *generalizable* patterns. For instance, a complex equation using many characteristics of each student to predict law school GPA could fit last year's data splendidly, but it might be idiosyncratic to the "training" data. In other words, it might not generalize to next year's data. This is known as overfitting. It can occur with traditional statistical techniques but is a greater risk with the high-dimensional (many variables) algorithmic approaches

219. For reviews of ML studies identifying frequent mistakes or departures from good practice, see Sayash Kapoor & Arvind Narayanan, *Leakage and the Reproducibility Crisis in Machine-learning-based Science*, 4 Patterns 100804 (2023), https://doi.org/10.1016/j.patter.2023.100804; Michael Roberts et al., *Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans*, 3 Nature Machine Intelligence 199 (2021); Laure Wynants et al., *Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal*, 369 Brit. Med. J. m1328 (2020), https://doi.org/10.1136/bmj.m1328. The extent to which published studies using ML-methods produce results that are superior to more traditional statistical methods is an active area of research. *See, e.g.,* Mohammad Ziaul Islam Chowdhury et al., *Prediction of Hypertension Using Traditional Regression and Machine Learning Models: A Systematic Review and Meta-Analysis*, 17 PLoS ONE e0266334 (2022), https://doi.org/10.1371/journal.pone.0266334; Xuan Song et al., *Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis*, 151 Int'l J. Med. Informatics 104484 (2021), https://doi.org/10.1016/j.ijmedinf.2021.104484.

220. The choice of statistics to indicate the accuracy of a predictor or classifier is another important topic.

221. Daniel Bone et al., *Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises*, 45 J. Autism & Developmental Disorders 1121 (2015), https://doi.org/10.1007/s10803-014-2268-6.

that characterize modern machine learning. "Internal validation" can be achieved by holding out some data in the training set as the algorithm is developed and measuring how well the algorithm predicts or classifies those test data.[222] But such internal validation is not sufficient. "External validation" requires testing the algorithm on data from a different source, and results from those tests should be reported. Furthermore, even if external validation has been performed at one point in time, confidence in continued predictive accuracy requires ongoing efforts to ensure that the relationships among variables in the data are not changing over time (or to revise the equations or algorithms if they are).

Is the Predictor or Classifier Too Opaque?

Finally, ML algorithms can be difficult to interpret, especially when the patterns that are "learned" may not be represented in easily decoded forms. This can make it hard to tell when the data environment is changing in ways that will lead to poorer performance. It also can mask the fact that the algorithm has access to features that should not be part of the data. For example, when researchers rushed to apply ML methods to COVID-19 diagnosis from chest X-rays and CT scans, many of them unwittingly used a dataset that contained chest scans of children who did not have COVID as their examples of what non-COVID cases looked like. As a result, the ML diagnoses were relying on features identifying children rather than COVID.[223] Concerns such as these are prompting the development of methods for gaining insight into how opaque ML systems are functioning[224] and to lists of criteria for judging when ML algorithms are most likely to be trustworthy.[225]

222. This is a very rough description of cross-validation with data available when developing the algorithm. For discussion of data-splitting and resampling techniques, see, for example, Max Kuhn & Kjell Johnson, *Applied Predictive Modeling* 61–92 (2013).

223. *Id.*

224. Berk, *supra* note 211, at 231.

225. *E.g.*, Kapoor & Narayanan, *supra* note 219; Russell A. Poldrack et al., *Establishment of Best Practices for Evidence for Prediction: A Review*, 77 JAMA Psychiatry 534 (2020), https://doi.org/10.1001/jamapsychiatry.2019.3671; Roberts et al., *supra* note 219.

Appendix: Conditional Probability and Bayes' Rule

What Do Probabilities Apply To?

The mathematical theory of probability consists of theorems derived from axioms and definitions. Mathematical reasoning is seldom controversial, but there may be disagreement as to how the theory should be applied. For example, statisticians may differ on the interpretation of data in specific applications. Moreover, there are two main schools of thought about the foundations of statistics: frequentist and Bayesian (also called objectivist and subjectivist).[226]

Frequentists see probabilities as empirical facts. When a fair coin is tossed, the probability of heads is taken to be 1/2; this value is justified by the argument that if the experiment is repeated a large number of times, the coin will land heads about one-half the time. If a fair die is rolled, the probability of getting an ace (one spot) is 1/6. If the die is rolled many times, an ace will turn up about one-sixth of the time.[227] Generally, if a chance experiment can be repeated, the relative frequency of an event approaches (in the long run) its probability. By contrast, a Bayesian considers probabilities as representing not facts but degrees of belief: in whole or in part, probabilities are subjective.[228]

226. The extent to which statisticians using Bayesian methods are overtly subjective in their analyses varies. "Objective Bayesians" use Bayes' rule without eliciting prior probabilities from subjective beliefs. One strategy is to use preliminary data to estimate the prior probabilities and then apply Bayes' rule to that empirical distribution. This "empirical Bayes" procedure avoids the charge of subjectivism at the cost of departing from a fully Bayesian framework. With ample data, it can be effective, and the estimates or inferences can be understood in frequentist terms. Another "objective" approach is to use "noninformative" priors that are supposed to be independent of all data and prior beliefs. However, the choice of such priors can be questioned, and the approach has been contested by frequentists and subjective Bayesians. *E.g.,* Joseph B. Kadane, *Is "Objective Bayesian Analysis" Objective, Bayesian, or Wise?*, 1 Bayesian Analysis 433 (2006), https://perma.cc/CSN9-35G5; Jon Williamson, *Philosophies of Probability, in* Philosophy of Mathematics 493 (Andrew Irvine ed., 2009) (discussing the challenges to objective Bayesianism).

227. Probabilities may be estimated from relative frequencies, but probability itself is a subtler idea. For example, suppose a computer prints out a sequence of ten letters H and T (for heads and tails), which alternate between the two possibilities H and T as follows: H T H T H T H T H T. The relative frequency of heads is 5/10 or 50%, but it is not at all obvious that the chance of an H at the next position is 50%. There are difficulties in both the subjectivist and objectivist positions. *See* Freedman, *supra* note 167.

228. "Subjective" is not necessarily the same as arbitrary. The degrees of belief must obey the same mathematical rules (axioms and theorems) as relative frequencies do, and the personal choices for particular numbers can be subjected to interpersonal standards for acceptance by other individuals.

What Are Conditional Probabilities?

Conditional probability is the probability of one event given that another has occurred. For example, suppose a fair coin is tossed twice. One event is that the coin will land heads up both times (HH). Another event is that at least one H will be seen. Before the coin is tossed, there are four possible, equally likely, outcomes: HH, HT, TH, TT. So the probability of HH is 1/4. However, if we know that at least one head has been obtained, then we can rule out two tails TT. In other words, given that at least one H has been obtained, the conditional probability of TT is 0, and the first three outcomes have conditional probability 1/3 each. In particular, the conditional probability of HH is 1/3. This is usually written as P(HH|at least one H) = 1/3. More generally, the probability of an event C is denoted P(C); the conditional probability of D given C is written as P(D|C).

Two events C and D are independent if the conditional probability of D given that C occurs is equal to the conditional probability of D given that C does not occur. Using ~C to denote the event that C does not occur, C and D are independent if P(D|C) = P(D|~C). If C and D are independent, then the probability that both occur is equal to the product of the probabilities:

$$P(C \text{ and } D) = P(C) \times P(D) \tag{3}$$

This is the multiplication rule (or product rule) for independent events. If events are dependent, then conditional probabilities must be used:

$$P(C \text{ and } D) = P(C) \times P(D|C) \tag{4}$$

This is the multiplication rule for dependent events. Statisticians caution against, but sometimes succumb to, using the multiplication rule for independent events when the events are dependent.[229]

What Is Bayes' Rule?

If we use probabilities to describe uncertainty regarding hypotheses as well as to describe events, we can use the formulas for conditional probabilities to obtain an equation known as Bayes' rule that has been proposed to guide legal factfinding. If

229. *E.g.*, William Kruskal, *Miracles and Statistics: The Casual Assumption of Independence*, 83 J. Am. Stat. Ass'n 929 (1988), https://doi.org/10.2307/2290117. A famous case of multiplication without apparent attention to dependence is *People v. Collins*, 438 P.2d 33 (Cal. 1968). For more examples, see *infra* note 238.

two mutually exclusive hypotheses H_0 and H_1 describe all the ways that an event A could occur,[230] it is easy to show that[231]

$$\mathrm{P}\left(H_1 \mid A\right) = \frac{\mathrm{P}\left(A \mid H_1\right)\mathrm{P}\left(H_1\right)}{\mathrm{P}\left(A \mid H_0\right)\mathrm{P}\left(H_0\right) + \mathrm{P}\left(A \mid H_1\right)\mathrm{P}\left(H_1\right)}. \tag{5}$$

This is one way of expressing Bayes' rule. It yields the conditional probability of hypothesis H_1 given that event A has occurred.

For a stylized example in a criminal case, suppose that blood from a single source is found at the scene of a crime and that the defendant in the case has type A blood. It is natural to have one hypothesis H_0 propose that the blood came from a person other than the defendant; the other hypothesis H_1 is that the blood came from the defendant; the event A is the event that blood from the crime scene is found to be type A. Then $\mathrm{P}(H_0)$ is the "prior probability" of H_0, based on subjective judgment of the other evidence and background information in the case, while $\mathrm{P}(H_0 \mid A)$ is the "posterior probability"—updated from the prior probability using the data on the blood.

Type A blood occurs in 42% of the population. Assuming that the laboratory's findings are always correct, $\mathrm{P}(A \mid H_0) = 0.42$.[232] Because the defendant has type A blood, if he is the source, event A will occur: $\mathrm{P}(A \mid H_1) = 1$. Suppose the

230. "Mutually exclusive" means that either one hypothesis or the other—but not both—is true.

231. We use the multiplication rule (4) to find that

$$\mathrm{P}(A \text{ and } H_0) = \mathrm{P}(A \mid H_0)\ \mathrm{P}(H_0) \tag{6}$$

and

$$\mathrm{P}(A \text{ and } H_1) = \mathrm{P}(A \mid H_1)\ \mathrm{P}(H_1). \tag{7}$$

Moreover, if A can occur only when either H_0 or H_1 is true, then

$$\mathrm{P}(A) = \mathrm{P}(A \text{ and } H_0) + \mathrm{P}(A \text{ and } H_1). \tag{8}$$

The multiplication rule (4) also shows that

$$\mathrm{P}(H_1 \mid A) = \mathrm{P}(A \text{ and } H_1)\ /\ \mathrm{P}(A). \tag{9}$$

Using (7) to evaluate $\mathrm{P}(A \text{ and } H_1)$ in the numerator of (9), and (6), (7), and (8) to evaluate $\mathrm{P}(A)$ in the denominator gives equation (5) in the text.

232. Not all statisticians would accept the identification of a population frequency with $\mathrm{P}(A \mid H_0)$. Indeed, H_0 has been translated into a hypothesis that the true donor has been selected from the population at random (i.e., in a manner that is uncorrelated with blood type). This step needs justification.

prior probabilities are $P(H_0) = P(H_1) = 0.5$. According to (5), the posterior probability that the blood is from the defendant is

$$P(H_1 \mid A) = \frac{1 \times 0.5}{(0.42 \times 0.5) + (1 \times 0.5)} = 0.70 \tag{10}$$

Thus, the data increase the probability that the blood is the defendant's. The probability went up from the prior value of $P(H_1) = 0.50$ to the posterior value of $P(H_1 \mid A) = 0.70$.

Equation (5) can be rewritten as follows:

$$\frac{P(H_1 \mid A)}{P(H_0 \mid A)} = \frac{P(A \mid H_1)}{P(A \mid H_0)} \times \frac{P(H_1)}{P(H_0)}. \tag{11}$$

This form gives us a convenient way to express the idea that the data change the chances in favor of H_1 as opposed to H_0. The ratio of the two prior probabilities is the odds in favor of H_1 as opposed to H_0.[233] In terms of odds, the version of Bayes' rule in (5) becomes[234]

$$\textit{Posterior odds} = \textit{Likelihood ratio} \times \textit{Prior odds} \tag{12}$$

where the "likelihood ratio" is the probability of the data (the crime-scene blood is type A) under the hypotheses H_1 divided by the probability of the same data under the other hypothesis H_0. This ratio states how many times more (or less) probable the data are under H_1 as opposed to H_0. Here, it is $1/0.42 = 2.31$. We would expect the data (type A blood) more than twice as often for the defendant than for a randomly selected individual; to that extent, it supports the hypothesis that the defendant is the source. It is more compatible with that hypothesis than the alternative being considered.

Used with Bayes' rule, this likelihood ratio means that the data increase the prior odds by a factor of a little more than 2, from 1:1 to about 2.31:1. Odds of 2.31 to 1 are the same as a probability of $2.31/(2.31 + 1) = 0.70$. Equations (5), (11), and (12) are all forms of the same fundamental relationship. In the context of Bayes' rule, the likelihood ratio is called a Bayes' factor.[235] It is the *change* in the

233. If the odds in favor of an event or a hypothesis are j to k, the probability is j divided by $j + k$. For example, favorable odds of seven to three (more compactly written as 7:3 or 7/3) indicate a probability of $7/10 = 0.70$.

234. Unlike equation (5), equations (11) and (12) hold even when other hypotheses could account for A.

235. *See* Kaye, *supra* note 172; *cf.* Stern, *supra* note 110 (summarizing frequentist, likelihood, and Bayesian frameworks for statistical inference). The likelihood ratio for binary test results or decisions, such as those in Table 1, is the true positive proportion (sensitivity) divided by the false positive probability (1—specificity). Low sensitivity degrades the value of a positive test result,

odds, which can be quite different from—and should not be confused with—the posterior odds.[236]

The same ideas can be applied more broadly to the various statistical models that have been illustrated in this reference guide. There are Bayesian approaches to fitting regression or other statistical models to make predictions and to obtain estimates of population parameters.[237] But whatever mode of statistical inference is employed, assessing probabilities, conditional probabilities, and independence is not entirely straightforward. Inquiry into the basis for expert judgment may be useful, and casual assumptions about independence should be questioned.[238]

which is why the false-positive probability is not a complete measure of the probative value of a positive finding.

236. For cases in which this mistake has been made, *see* David H. Kaye, *Reference Guide on Human DNA Identification Evidence*, "Presentation of LRs" section, in this manual. The numerical disparity between the posterior odds and the Bayes' factor will depend on its magnitude of the Bayes' factor and on the prior odds. For example, the two are numerically equal when the prior odds are 1:1 (a prior probability of 50%). But if we insert prior odds of 1:10 into (12), the posterior odds become 2.31:10, or about 7:30. For these lower prior odds, the posterior probability is about 7/37 = 0.19. Even though the blood-type evidence supports the same-source hypothesis over the different-source hypothesis by a factor of more than two, the same-source hypothesis remains improbable.

237. Modern treatments of Bayesian methods include Donald A. Berry, *Statistics: A Bayesian Perspective* (1995); Andrew Gelman et al., *Bayesian Data Analysis* (3d ed. 2013); Jeff Gill, *Bayesian Methods: A Social and Behavioral Sciences Approach* (3d ed. 2015); John K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2d ed. 2015); Richard McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2d ed. 2020).

238. For problematic assumptions of independence in litigation, see, for example, *Wilson v. State*, 803 A.2d 1034 (Md. 2002) (error to admit multiplied probabilities in a case involving two deaths of infants in same family); 1 McCormick, *supra* note 1, § 210; see also *supra* note 36 (on census litigation).

Glossary of Terms

The following definitions are adapted from a variety of sources, including Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001), David A. Freedman et al., *Statistics* (4th ed. 2007), and David Spiegelhalter, *The Art of Statistics: How to Learn from Data* (2019).

absolute value. Size, neglecting sign. The absolute value of +2.7 is 2.7; so is the absolute value of −2.7.

adjust for. See control for.

alpha (α). A symbol often used to denote the probability of a Type I error. See Type I error; size. Compare beta.

alternative hypothesis. A statistical hypothesis that is contrasted with the null hypothesis in a significance test. See statistical hypothesis; significance test.

area sample. A probability sample in which the sampling frame is a list of geographical areas. That is, the researchers make a list of areas, choose some at random, and interview people in the selected areas. This is a cost-effective way to draw a sample of people. See probability sample; sampling frame.

arithmetic mean. See mean.

average. See mean. "Average" often is used generically, to refer to a single representative or central value, such as the mean, median, or mode for a set of numbers.

Bayes factor. A ratio that indicates how much the data changes the prior odds. See Bayes' rule.

Bayes' rule. In its simplest form, an equation involving conditional probabilities that relates a "prior probability" known or estimated before collecting certain data to a "posterior probability" that reflects the impact of the data on the prior probability.

Bayesian inference. In Bayesian statistical inference, "the prior" expresses degrees of belief about various hypotheses or parameters before data are collected. Data are then collected according to some statistical model; at least, the model represents the investigator's beliefs. Bayes' rule combines the prior probability with the data to yield the posterior probability, which expresses the investigator's beliefs about the hypotheses or parameters, given the data. See Appendix. Compare frequentist.

beta (β). A symbol sometimes used to denote power, and sometimes to denote the probability of a Type II error. See Type II error; power. Compare alpha.

between-observer variability. Differences that occur when two or more observers measure the same thing. Compare within-observer variability.

bias. Also called systematic error. A systematic tendency for an estimate to be too high or too low. An estimate is unbiased if the bias is zero. (Bias does not mean prejudice, partiality, or discriminatory intent.) See nonsampling error. Compare sampling error.

bias-variance trade-off. Some estimators computed from a sample have less variability across samples than an unbiased estimator, making it reasonable to accept some bias in order to increase precision (reduce the standard error of the estimate). Similarly, when fitting a model for prediction, increasing complexity will eventually lead to a model that has less bias, in the sense that it has greater potential to adapt to details of the underlying process, but more variance, since there is not enough data to be confident about the parameters in the model. These elements need to be traded off to avoid over-fitting.

big data. A huge volume of data, often from a variety of sources such as images, social media accounts, or transactions, having a high velocity of acquisition and a possible lack of veracity due to its routine collection.

bin. A class interval in a histogram. See class interval; histogram.

binary variable. A variable that has only two possible values (e.g., sex). Called a dummy variable when the two possible values are 0 and 1.

binomial distribution. A distribution for the number of occurrences in repeated, independent "trials" where the probabilities are fixed. For example, the number of heads in 100 tosses of a coin follows a binomial distribution. When the probability is not too close to 0 or 1 and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See normal distribution; Poisson distribution.

blind. See double-blind experiment.

bootstrap. Also called resampling; Monte Carlo method. A procedure for estimating sampling error by constructing a simulated population on the basis of the sample, then repeatedly drawing samples from the simulated population.

categorical data; categorical variable. See qualitative variable. Compare quantitative variable.

central limit theorem. Mathematical theorem showing that under suitable conditions, the probability histogram for a sum (or average or rate) will follow the normal curve. See histogram; normal curve.

chance error. See random error; sampling error.

chi-squared (χ^2). The chi-squared statistic measures the distance between the data and expected values computed from a statistical model. If the chi-squared statistic is too large to explain by chance, the data contradict the model. The definition of "large" depends on the context. See statistical hypothesis; significance test.

class interval. Also, bin. The base of a rectangle in a histogram; the area of the rectangle shows the frequency or relative frequency of observations in the class interval. See histogram.

cluster sample. A type of random sample. For example, investigators might take households at random, then interview all people in the selected households. This is a cluster sample of people: A cluster consists of all the people in a selected household. Generally, clustering reduces the cost of interviewing. See multistage cluster sample.

coefficient of determination. A statistic (more commonly known as *R*-squared) that describes how well a regression equation fits the data. See *R*-squared.

coefficient of variation. A statistic that measures spread relative to the mean: SD/mean, or SE/expected value. See expected value; mean; standard deviation; standard error.

collinearity. See multicollinearity.

conditional probability. The probability that one event will occur given that another has occurred.

confidence coefficient. See confidence interval.

confidence interval. An estimate, expressed as a range, for a parameter. For estimates such as averages or rates computed from large samples, a 95% confidence interval is the range from about two standard errors below to two standard errors above the estimate. Intervals obtained this way cover the true value about 95% of the time, and 95% is the confidence level or the confidence coefficient. See central limit theorem; standard error.

confidence level. See confidence interval.

confounding variable; confounder. A confounder is correlated with the independent variable and the dependent variable. An association between the dependent and independent variables in an observational study may not be causal, but may instead be due to confounding. See controlled experiment; observational study.

consistent estimator. An estimator that tends to become more and more accurate as the sample size grows. Inconsistent estimators, which do not become more accurate as the sample gets larger, are frowned upon by statisticians.

content validity. The extent to which a skills test is appropriate to its intended purpose, as evidenced by a set of questions that adequately reflect the domain being tested. See validity. Compare reliability.

continuous variable. A variable that has arbitrarily fine gradations, such as a person's height. Compare discrete variable.

control for. Statisticians may control for the effects of confounding variables in nonexperimental data by making comparisons for smaller and more

homogeneous groups of subjects, or by entering the confounders as explanatory variables in a regression model. To "adjust for" is perhaps a better phrase in the regression context, because in an observational study the confounding factors are not under experimental control; statistical adjustments are an imperfect substitute. See regression model.

control group. See controlled experiment.

controlled experiment. An experiment in which the investigators determine which subjects are put into the treatment group and which are put into the control group. Subjects in the treatment group are exposed by the investigators to some influence (the treatment); those in the control group are not so exposed. For example, in an experiment to evaluate a new drug, subjects in the treatment group are given the drug, and subjects in the control group are given some other therapy; the outcomes in the two groups are compared to see whether the new drug works. Randomization—that is, randomly assigning subjects to each group—is usually the best way to ensure that any observed difference between the two groups comes from the treatment rather than from preexisting differences. In many situations, a randomized controlled experiment is impractical, and investigators must then rely on observational studies. Compare observational study.

convenience sample. A nonrandom sample of units, also called a grab sample. Such samples are easy to take but may suffer from serious bias. Typically, mall samples are convenience samples.

correlation coefficient. A number between –1 and 1 that indicates the extent of the linear association between two variables. Often, the correlation coefficient is abbreviated as *r*.

covariance. A quantity that describes the statistical interrelationship of two variables. Compare correlation coefficient; standard error; variance.

covariate. A variable that is related to other variables of primary interest in a study; a measured confounder; a statistical control in a regression equation.

credible interval. A range of values obtained through a Bayesian analysis that contains a parameter with a specified degree of belief.

criterion. The variable against which an examination or other selection procedure is validated. See validity.

data. Observations or measurements, usually of units in a sample taken from a larger population.

data science. The study and application of techniques for deriving insights from data, including constructing algorithms for prediction. Traditional statistical science forms part of data science, which also includes a strong element of computer science and data management.

degrees of freedom. See *t*-test.

dependence. Two events are dependent when the probability of one is affected by the occurrence or non-occurrence of the other. Compare independent; dependent variable.

dependent variable. Also called outcome variable. Compare independent variable.

descriptive statistics. Like the mean or standard deviation, used to summarize data.

differential validity. Differences in validity across different groups of subjects. See validity.

discrete variable. A variable that has only a small number of possible values, such as the number of automobiles owned by a household. Compare continuous variable.

distribution. See frequency distribution; probability distribution; sampling distribution.

disturbance term. A synonym for error term.

double-blind experiment. An experiment with human subjects in which neither the diagnosticians nor the subjects know who is in the treatment group or the control group. This is accomplished by giving a placebo treatment to patients in the control group. In a single-blind experiment, the patients do not know whether they are in treatment or control; the diagnosticians have this information.

dummy variable. See indicator variable.

econometrics. Statistical study of economic issues.

epidemiology. Statistical study of disease or injury in human populations.

error term. The part of a statistical model that describes random error or chance variation, i.e., the impact of chance factors unrelated to variables in the model. In econometrics, the error term is called a disturbance term.

estimator. A sample statistic used to estimate the value of a population parameter. For example, the sample average commonly is used to estimate the population average. The term "estimator" connotes a statistical procedure, whereas an "estimate" connotes a particular numerical result.

expected value. The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities of the values. For example, the spots that turn up when a pair of dice are tossed is a random variable, and the expected value is

$$\frac{1}{36}\times 2+\frac{2}{36}\times 3+\frac{3}{36}\times 4+\frac{4}{36}\times 5+\frac{5}{36}\times 6+\frac{6}{36}\times 7$$
$$+\frac{5}{36}\times 8+\frac{4}{36}\times 9+\frac{3}{36}\times 10+\frac{2}{36}\times 11+\frac{1}{36}\times 12$$

which is equal to 7.

experiment. See controlled experiment; randomized controlled experiment. Compare observational study.

explanatory variable. See independent variable; regression model.

external validity. See validity.

factors. See independent variable.

false negative. In a statistical analysis, the decision not to reject the null hypothesis when it is in fact false. In studies of forensic-science identification procedures, it is clearer to use terminology such as "false identification" or "false exclusion" to specify the type of incorrect decision being addressed.

false positive. In a statistical analysis, the decision to reject the null hypothesis when it is in fact true. In studies of forensic-science identification procedures, it is clearer to use terminology such as "false identification" or "false exclusion" to specify the type of incorrect decision being addressed.

Fisher's exact test. A statistical test for comparing two sample proportions. For example, take the proportions of white and Black employees getting a promotion. An investigator may wish to test the null hypothesis that promotion does not depend on race. Fisher's exact test is one way to arrive at a *p*-value. The calculation is based on the hypergeometric distribution. See hypergeometric distribution; *p*-value; significance test; statistical hypothesis.

fitted value. See residual.

fixed significance level. Also alpha; size. A preset level, such as 5% or 1%; if the *p*-value of a test falls below this level, the result is deemed statistically significant. See significance test. Compare observed significance level; *p*-value.

frequency; relative frequency. Frequency is the number of times that something occurs; relative frequency is the number of occurrences, relative to a total. For example, if a coin is tossed 1,000 times and lands heads 517 times, the frequency of heads is 517; the relative frequency is 0.517, or 51.7%.

frequency distribution. Shows how often specified values occur in a dataset.

frequentist. Describes statisticians who view probabilities as objective properties of a system that can be measured or estimated and who use statistical procedures justified by their performance in repeated samples from the population of interest. Compare Bayesian inference. See Appendix.

Gaussian distribution. A synonym for normal distribution.

general linear model. Expresses the dependent variable as a linear combination of the independent variables plus an error term whose components may be dependent and have differing variances. See error term; linear combination; variance. Compare regression model.

grab sample. See convenience sample.

heteroscedastic. See scatter diagram.

highly significant. See *p*-value; practical significance; significance test.

histogram. A plot showing how observed values fall within specified intervals, called bins or class intervals. Generally, matters are arranged so that the area under the histogram for each class interval gives the frequency or relative frequency of data in that interval. With a probability histogram, the area gives the chance of observing a value that falls in the interval.

homoscedastic. See scatter diagram.

hypergeometric distribution. Suppose a sample is drawn at random, without replacement, from a finite population. How many times will items of a certain type come into the sample? The hypergeometric distribution gives the probabilities. Compare Fisher's exact test.

hypothesis. See alternative hypothesis; null hypothesis; one-sided hypothesis; significance test; statistical hypothesis; two-sided hypothesis.

hypothesis test. Also, null hypothesis test; statistical test; test of significance. Involves formulating a statistical hypothesis (the null hypothesis) and an alternative hypothesis; devising a test statistic; and defining a critical region in which the test statistic prompts the rejection of the null hypothesis. The critical region (also called a rejection region) is constructed so that the probability of rejecting the null hypothesis—if the hypothesis is true—is no larger than some preestablished value (often $\alpha = 5\%$). Hypothesis tests often are performed by computing the *p*-value of the observed test statistic and comparing it to the preset value for rejection. See also significance test.

identically distributed. Random variables are identically distributed when they have the same probability distribution. For example, consider a box of numbered tickets. Draw tickets at random with replacement from the box. The draws will be independent and identically distributed.

independence. Also, statistical independence. Events are independent when the probability of one is unaffected by the occurrence or nonoccurrence of the other. Compare conditional probability; dependence; independent variable; dependent variable.

independent variable. Independent variables (also called explanatory variables, predictors, or risk factors) represent the causes and potential confounders in a statistical study of causation; the dependent variable represents the outcome or effect. In an observational study, independent variables may be used to divide the population into smaller and more homogenous groups ("stratification"). In a regression model, the independent variables are used to predict the dependent variable. For example, the unemployment rate has been used as the independent variable in a model for predicting the crime rate; the unemployment rate is the independent variable in this model, and the crime rate is the

dependent variable. The distinction between independent and dependent variables is unrelated to statistical independence. See regression model. Compare dependent variable; dependence; independence.

indicator variable. Variable created to assign numerical values to categorical or nominal data. The most common approach is to use a dummy variable that takes only the values 0 or 1 to distinguish one group of interest from another. See binary variable; regression model.

internal validity. See validity.

interquartile range. Difference between 25th and 75th percentile. See percentile.

interval estimate. A confidence interval, or an estimate coupled with a standard error. See confidence interval; standard error. Compare point estimate.

least squares. See least squares estimator; regression model.

least squares estimator. An estimator that is computed by minimizing the sum of the squared residuals. See residual.

level. The level of a significance test is denoted alpha (α). See alpha; fixed significance level; observed significance level; *p*-value; significance test.

likelihood. Colloquially, the chance that something will happen. In statistical models, the probability of observing a specific value of an observable quantity; may depend on the values of parameters.

likelihood ratio. A ratio of the probability of observing a specific value of an observable quantity under different hypotheses or parameter values.

linear combination. To obtain a linear combination of two variables, multiply the first variable by some constant, multiply the second variable by another constant, and add the two products. For example, $2u + 3v$ is a linear combination of u and v.

linear regression. A statistical model relating a continuous outcome or response to independent variable(s) or predictor(s). The dependent variable is modeled as a linear combination of the independent ones (or a transformed version of them) plus an error term whose values are randomly distributed in some fashion.

list sample. See systematic sample.

logistic regression. A statistical model relating a binary outcome or response to independent variable(s) or predictor(s). The logarithm of the odds of an event occurring is modelled as a linear combination of the independent variables plus a randomly distributed error term.

loss function. Statisticians may evaluate estimators according to a mathematical formula involving the errors—that is, differences between actual values and estimated values. The "loss" may be the total of the squared errors, or the total of the absolute errors, etc. Loss functions seldom quantify real losses,

but may be useful summary statistics and may prompt the construction of useful statistical procedures. Compare risk.

lurking variable. See confounding variable.

Markov chain. A random process describing a sequence of variables or events in which the probability of each variable or event depends only on the values of the immediately preceding variables or events.

mean. Also, the average; the expected value of a random variable. The mean gives a way to find the center of a batch of numbers: Add the numbers and divide by how many there are. Weights may be employed, as in "weighted mean" or "weighted average." See random variable. Compare median; mode.

measurement validity. See validity. Compare reliability.

median. The median, like the mean, is a way to find the center of a batch of numbers. The median is the 50th percentile. Half the numbers are larger, and half are smaller. (To be very precise: at least half the numbers are greater than or equal to the median; at least half the numbers are less than or equal to the median; for small datasets, the median may not be uniquely defined.) Compare mean; mode; percentile.

meta-analysis. Attempts to combine information from all studies on a certain topic. For example, in the epidemiological context, a meta-analysis may attempt to provide a summary odds ratio and confidence interval for the effect of a certain exposure on a certain disease.

metrology. The scientific study of measurement.

mode. The most common number in a batch of numbers. Compare mean; median.

model. See probability model; regression model; statistical model.

Monte Carlo method. Relies on random sampling to estimate quantities of interest. The quantities of interest may be stochastic or deterministic. Monte Carlo methods give approximate solutions to many mathematical and statistical problems for which no simple analytical solution is available.

multicollinearity. Also, collinearity. The existence of correlations among the independent variables in a regression model. See independent variable; regression model.

multiple comparison. Making several statistical tests on the same dataset. Multiple comparisons complicate the interpretation of a p-value. For example, if twenty divisions of a company are examined, and one division is found to have a disparity significant at the 5% level, the result is not surprising; indeed, it would be expected under the null hypothesis. Compare p-value; significance test; statistical hypothesis.

multiple correlation coefficient. A number that indicates the extent to which one variable can be predicted as a linear combination of other variables. Its

magnitude is the square root of *R*-squared. See linear combination; *R*-squared; regression model. Compare correlation coefficient.

multiple regression. A regression equation that includes two or more independent variables. See regression model. Compare simple regression.

multistage cluster sample. A probability sample drawn in stages, usually after stratification; the last stage will involve drawing a cluster. See cluster sample; probability sample; stratified random sample.

multivariate methods. Methods for fitting models with multiple variables; in statistics, multiple response variables; in other fields, multiple explanatory variables. See regression model.

natural experiment. An observational study in which treatment and control groups have been formed by some natural development, but the assignment of subjects to groups is akin to randomization. See observational study. Compare controlled experiment.

nonparametric method. A method for testing a hypothesis or obtaining an interval that does not require knowledge of the form of the underlying parent population; also called a distribution-free method.

nonresponse bias. Systematic error created by differences between respondents and nonrespondents. If the nonresponse rate is high, this bias may be severe.

nonsampling error. A catch-all term for sources of error in a survey, other than sampling error. Nonsampling errors cause bias. One example is selection bias: The sample is drawn in a way that tends to exclude certain subgroups in the population. A second example is nonresponse bias: People who do not respond to a survey are usually different from respondents. A final example: Response bias arises if the interviewer uses a loaded question (or for other reasons).

normal distribution. Also, Gaussian distribution. Describes the probability density of certain continuous variables. The family of normal curves has two parameters: the mean and the standard deviation. The equation for the normal probability density function is given in the section "The normal curve and large samples." Terminology notwithstanding, there need be nothing wrong with a distribution that differs from normal.

null hypothesis. For example, a hypothesis that there is no difference between two groups from which samples are drawn. See significance test; statistical hypothesis. Compare alternative hypothesis.

observational study. A study in which subjects (or external forces) select themselves into groups; investigators then compare the outcomes for the different groups. For example, studies of smoking are generally observational. Subjects decide whether or not to smoke; the investigators compare the death rate for smokers to the death rate for nonsmokers. In an observational study, the groups may differ in important ways that the investigators do not notice; controlled

experiments minimize this problem. The critical distinction is that in a controlled experiment, the investigators intervene to manipulate the circumstances of the subjects; in an observational study, the investigators are passive observers. (Of course, running a good observational study is hard work, and may be quite useful.) Compare confounding variable; controlled experiment.

observed significance level. A synonym for *p*-value. See significance test. Compare fixed significance level.

odds. The probability that an event will occur divided by the probability that it will not. For example, if the chance of rain tomorrow is 2/3, then the odds of rain are (2/3)/(1/3) = 2/1, or 2 to 1; the odds against rain are 1 to 2.

odds ratio. A measure of association, often used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then the odds of the disease in the exposed group are 10/90 = 1/9, compared to 5/95 = 1/19 in the unexposed group. The odds ratio is (1/9)/(1/19) = 19/9 = 2.1. An odds ratio of 1 indicates no association. Compare relative risk.

one-sided hypothesis; one-tailed hypothesis. Excludes the possibility that a parameter could be, for example, less than the value asserted in the null hypothesis. A one-sided hypothesis leads to a one-sided (or one-tailed) test. See significance test; statistical hypothesis. Compare two-sided hypothesis.

one-sided test; one-tailed test. See one-sided hypothesis.

overfitting. Occurs when a statistical model, especially one with many free parameters, fits the data used to estimate the model extremely well while generalizing poorly to other data for which it is intended.

outcome variable. See dependent variable.

outlier. An observation that is far removed from the bulk of the data. Outliers may indicate faulty measurements, and they may exert undue influence on summary statistics, such as the mean or the correlation coefficient.

***p*-value.** Result from a statistical test. The probability of getting, just by chance, a test statistic as large as or larger than the observed value. Large *p*-values are consistent with the null hypothesis; small *p*-values undermine the null hypothesis. However, *p* does not give the probability that the null hypothesis is true. The *p*-value is a useful measure of compatibility with the null hypothesis. If *p* is smaller than 5%, the result is often said to be statistically significant. If *p* is smaller than 1%, the result may be described as highly significant. Strict reliance on such cutoffs is not recommended as a best practice. The *p*-value is also called the observed significance level. See significance test; statistical hypothesis.

parameter. A numerical characteristic of a population or a model. See probability model.

percentile. To get the percentiles of a dataset, array the data from the smallest value to the largest. As an example of a percentile, consider the 90th percentile: 90% of the values fall below the 90th percentile and 10% are above. (To be very precise: at least 90% of the data are at the 90th percentile or below; at least 10% of the data are at the 90th percentile or above.) The 50th percentile is the median: 50% of the values fall below the median, and 50% are above.

placebo. See double-blind experiment.

point estimate. An estimate of the value of a quantity expressed as a single number. See estimator. Compare confidence interval; interval estimate.

Poisson distribution. A limiting case of the binomial distribution, when the number of trials is large and the common probability is small. The parameter of the approximating Poisson distribution is the number of trials times the common probability, which is the expected number of events. When this number is large, the Poisson distribution may be approximated by a normal distribution.

population. Also, universe. All the units of interest to the researcher. Compare sample; sampling frame.

population size. Also, size of population. Number of units in the population.

posterior probability. See Bayes' rule.

power. The probability that a statistical test will reject the null hypothesis. To compute power, one has to fix the size of the test and specify parameter values outside the range given by the null hypothesis. A powerful test has a good chance of detecting an effect when there is an effect to be detected. See beta; significance test. Compare alpha; size; *p*-value.

practical significance. Substantive importance. Statistical significance does not necessarily establish practical significance. With large samples, small differences can be statistically significant. See significance test.

practice effects. Changes in test scores that result from taking the same test twice in succession, or taking two similar tests one after the other.

predicted value. See residual.

predictive validity. A skills test has predictive validity to the extent that test scores are well correlated with later performance, or more generally with outcomes that the test is intended to predict. See validity. Compare reliability.

predictor. See independent variable.

prior probability. See Bayes' rule.

probability. Chance, on a scale from 0 to 1. Impossibility is represented by 0, certainty by 1. Equivalently, chances may be quoted in percent; 100% corresponds to 1, 5% corresponds to .05, and so forth.

probability density. Describes the probability distribution of a continuous random variable. The chance that the random variable falls in an interval equals the area below the density and above the interval. (However, not all random variables have densities.) See probability distribution; random variable.

probability distribution. Gives probabilities for possible values or ranges of values of a random variable. Often, the distribution is described in terms of a density. See probability density.

probability histogram. See histogram.

probability model. Relates probabilities of outcomes to parameters; also, statistical model. The latter connotes unknown parameters.

probability sample. A sample drawn from a sampling frame by some objective chance mechanism; each unit has a known probability of being sampled. Such samples minimize selection bias but can be expensive to draw.

psychometrics. The study of psychological measurement and testing.

qualitative variable; quantitative variable. A qualitative variable describes nonquantitative features of subjects in a study (e.g., marital status: never married, married, widowed, divorced, separated). A quantitative variable describes numerical features of the subjects (e.g., height, weight, income). This is not a hard-and-fast distinction, because qualitative features may be given numerical codes, as with a dummy variable. Quantitative variables may be classified as discrete or continuous. Concepts such as the mean and the standard deviation apply only to quantitative variables. Compare continuous variable; discrete variable; dummy variable. See variable.

quartile. The 25th or 75th percentile. See percentile. Compare median.

quasi-experiment. See natural experiment.

***R*-squared (R^2).** Measures how well a regression equation fits the data. *R*-squared varies between 0 (no fit) and 1 (perfect fit). *R*-squared does not measure the extent to which underlying assumptions are justified. See regression model. Compare multiple correlation coefficient; standard error of regression.

random error. Sources of error that are random in their effect, like draws made at random from a box. These are reflected in the error term of a statistical model. Some authors refer to random error as chance error or sampling error. See regression model.

random variable. A variable whose possible values occur according to some probability mechanism. For example, if a pair of dice are thrown, the total number of spots is a random variable. The chance of two spots is 1/36, the chance of three spots is 2/36, and so forth; the most likely number is 7, with a chance of 6/36. The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities. The expected value need not be a possible value for the random variable.

randomization. See controlled experiment; randomized controlled experiment.

randomized controlled experiment. A controlled experiment in which subjects are placed into the treatment and control groups at random—as if by a lottery. See controlled experiment. Compare observational study.

range. The difference between the biggest and the smallest values in a batch of numbers.

rate. In an epidemiological study, the number of events, divided by the size of the population; often cross-classified by age and gender. For example, the death rate from heart disease among American men in 2020 was about two per thousand. Among women, the rate was about half that.

regression coefficient. The coefficient of a variable in a regression equation. See regression model.

regression diagnostics. Procedures intended to check whether the assumptions of a regression model are appropriate.

regression equation. See regression model.

regression line. The graph of a (simple) regression equation.

regression model. A regression model attempts to combine the values of certain variables (the independent or explanatory variables) in order to get expected values for another variable (the dependent variable). Sometimes, the phrase "regression model" refers to a probability model for the data; if no qualifications are made, the model will generally be linear, and errors will be assumed independent across observations, with common variance. The coefficients in the linear combination are called regression coefficients; these are parameters. At times, "regression model" refers to an equation ("the regression equation") estimated from data, typically by least squares.

relative frequency. See frequency.

relative risk. A measure of association used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then the disease occurs twice as frequently among the exposed people: The relative risk is 10% divided by 5% = 2. A relative risk of 1 indicates no association. Compare odds ratio.

reliability. The extent to which a measurement process gives the same results on repeated measurement of the same thing. See repeatability, reproducibility. Compare validity.

repeatability. A type of reliability. With a repeatable procedure, measurements by the same examiner using the same equipment at the same place and time should be close to one another. See within-observer variability.

representative sample. Not a well-defined technical term. A sample judged to fairly represent the population, or a sample drawn by a process likely to give

samples that fairly represent the population, for example, a large probability sample.

reproducibility. A type of reliability. With a reproducible procedure, measurements from different examiners often at different places and times also should be similar. See between-observer variability.

resampling. See bootstrap.

residual. The difference between an actual and a predicted value. The predicted value comes typically from a regression equation, and is better called the fitted value, because there is no real prediction going on. See regression model; independent variable.

response variable. See independent variable.

risk factor. See independent variable.

robust. A statistic or procedure that does not change much when data or assumptions are modified slightly.

sample. A set of units collected for study. Compare population.

sample size. Also, size of sample. The number of units in a sample.

sample weights. See stratified random sample.

sampling distribution. The distribution of the values of a statistic, over all possible samples from a population. For example, suppose a random sample is drawn. Some values of the sample mean are more likely; others are less likely. The sampling distribution specifies the chance that the sample mean will fall in one interval rather than another.

sampling error. A sample is part of a population. When a sample is used to estimate a numerical characteristic of the population, the estimate is likely to differ from the population value because the sample is not a perfect microcosm of the whole. If the estimate is unbiased, the difference between the estimate and the exact value is sampling error. More generally, estimate = true value + bias + sampling error. Sampling error is also called chance error or random error. See standard error. Compare bias; nonsampling error.

sampling frame. A list of units designed to represent the entire population as completely as possible. The sample is drawn from the frame.

sampling interval. See systematic sample.

scatter diagram. Also, scatterplot; scattergram. A graph showing the relationship between two variables in a study. Each dot represents one unit from the study, e.g., one subject. One variable is plotted along the horizontal axis, the other variable is plotted along the vertical axis. A scatter diagram is homoscedastic when the spread is more or less the same inside any vertical strip. If the spread changes from one strip to another, the diagram is heteroscedastic.

selection bias. Systematic error due to nonrandom selection of subjects for study.

sensitivity. The probability that a test for the presence of a condition will give a positive result given that the condition is present. Sensitivity is analogous to the power of a statistical test. Compare specificity.

sensitivity analysis. Analyzing data in different ways to see how results depend on methods or assumptions.

sign test. A statistical test based on counting and the binomial distribution. For example, a Finnish study of twins found 22 monozygotic twin pairs where 1 twin smoked, 1 did not, and at least 1 of the twins had died. That sets up a race to death. In 17 cases, the smoker died first; in 5 cases, the nonsmoker died first. The null hypothesis is that smoking does not affect time to death, so the chances are 50–50 for the smoker to die first. On the null hypothesis, the chance that the smoker will win the race 17 or more times out of 22 is 8/1000. That is the p-value. The p-value can be computed from the binomial distribution.

significance level. See fixed significance level; p-value.

significance test. Testing for significance can refer to a hypothesis test performed by computing a p-value and declaring that the test statistic is significant if this p-value is less than or equal to some level (α). Significance testing also can denote just computing the statistic and its p-value to see whether the p-value is large or small; no formal test with a preset significance level is involved. The idea is to see whether the data conform to the predictions of the null hypothesis. Generally, a large test statistic goes with a small p-value; and small p-values would undermine the null hypothesis.

significant. See p-value; practical significance; significance test.

simple random sample. A simple random sample of size n from a population of size N is one in which each set of n units in the sampling frame has the same chance of being chosen as the sample. The investigators take a unit at random (as if by lottery), set it aside, take another at random from what is left, and so forth.

simple regression. A regression equation that includes only one independent variable. Compare multiple regression.

size. A synonym for alpha (α).

skip factor. See systematic sample.

specificity. The probability that a test for the presence of a condition will give a negative result given that the condition is not present. Specificity is analogous to $1-\alpha$, where α is the significance level of a statistical test. Compare sensitivity.

spurious correlation. When two variables are correlated, one is not necessarily the cause of the other. The vocabulary and shoe size of children in elementary school, for example, are correlated—but learning more words will not make the feet grow. Such noncausal correlations are said to be spurious. (Originally, the term seems to have been applied to the correlation between two rates with

the same denominator: Even if the numerators are unrelated, the common denominator will create some association.) Compare confounding variable.

standard deviation (SD). Indicates how far a typical element deviates from the average. For example, in round numbers, the average height of women age 18 and over in the United States is 5 feet 4 inches. However, few women are exactly average; most will deviate from average, at least by a little. The SD is sort of an average deviation from average. For the height distribution, the SD is 3 inches. The height of a typical woman is around 5 feet 4 inches, but is off that average value by something like 3 inches. For distributions that follow the normal curve, about 68% of the elements are in the range from 1 SD below the average to 1 SD above the average. Thus, about 68% of women have heights in the range 5 feet 1 inch to 5 feet 7 inches. Deviations from the average that exceed 3 or 4 SDs are extremely unusual. Many authors use standard deviation to also mean standard error. See standard error.

standard error (SE). Indicates the likely size of the sampling error in an estimate. Many authors use the term standard deviation instead of standard error. Compare expected value; standard deviation.

standard error of regression. Indicates how actual values differ (in some average sense) from the fitted values in a regression model. See regression model; residual. Compare *R*-squared.

standardization. See standardized variable.

standardized variable. A variable transformed to have mean zero and variance one. This involves two steps: (1) subtract the mean, (2) divide by the standard deviation.

statistic. A number that summarizes data. A statistic refers to a sample; a parameter or a true value refers to a population or a probability model.

statistical controls. Procedures that try to filter out the effects of confounding variables on non-experimental data, for example, by adjusting through statistical procedures such as stratification or multiple regression. See multiple regression; confounding variable; observational study. Compare controlled experiment.

statistical dependence. See dependence.

statistical hypothesis. Generally, a statement about parameters in a probability model for the data. The null hypothesis may assert that certain parameters have specified values or fall in specified ranges; the alternative hypothesis would specify other values or ranges. The null hypothesis is tested against the data with a test statistic; the null hypothesis may be rejected if there is a statistically significant difference between the data and the predictions of the null hypothesis. Typically, the investigator seeks to demonstrate the alternative hypothesis; the null hypothesis would explain the findings as a result of mere chance, and the investigator uses a significance or hypothesis test to rule out that possibility.

statistical independence. See independence.

statistical model. See probability model.

statistical significance. See *p*-value.

statistical test. See significance test.

stratified random sample. A type of probability sample. The researcher divides the population into relatively homogeneous groups called "strata" and draws a random sample separately from each stratum. Dividing the population into strata is called "stratification." Often the sampling fraction will vary from stratum to stratum. Then sampling weights should be used to extrapolate from the sample to the population. For example, if 1 unit in 10 is sampled from stratum A while 1 unit in 100 is sampled from stratum B, then each unit drawn from A counts as 10, and each unit drawn from B counts as 100. The first kind of unit has weight 10; the second has weight 100.

stratification. See independent variable; stratified random sample.

study validity. See validity.

subjectivist. See Bayesian inference.

systematic error. See bias.

systematic sample. Also, list sample. The elements of the population are numbered consecutively as 1, 2, 3, ... The investigators choose a starting point and a "sampling interval" or "skip factor" k. Then, every kth element is selected into the sample. If the starting point is 1 and $k = 10$, for example, the sample would consist of items 1, 11, 21, ... Sometimes the starting point is chosen at random from 1 to k: this is a random-start systematic sample.

***t*-statistic.** A test statistic, used to make the *t*-test. The *t*-statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested. With a large sample, a *t*-statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—the estimate is too many standard errors away from its expected value. See statistical hypothesis; significance test; *t*-test.

***t*-test.** A statistical test based on the *t*-statistic. Large *t*-statistics are beyond the usual range of sampling error. The *t*-test arises when testing the null hypothesis that the average of a population equals a given value when the population is known to be normally distributed. For small samples, the *t*-statistic follows Student's *t*-distribution (when the null hypothesis holds) rather than the normal curve; larger values of *t* are required to achieve significance with small samples. The relevant *t*-distribution depends on the number of degrees of freedom, which in this context equals the sample size minus one. A *t*-test is not appropriate for small samples drawn from a population that is not normal. See hypothesis test; *p*-value; significance test; statistical hypothesis.

test statistic. A statistic used to judge whether data conform to the null hypothesis. The parameters of a probability model determine expected values for the data; differences between expected values and observed values are measured by a test statistic. Such test statistics include the chi-squared statistic (χ^2) and the *t*-statistic. Generally, small values of the test statistic are consistent with the null hypothesis; large values lead to rejection. See hypothesis test; *p*-value; statistical hypothesis; *t*-statistic.

time series. A series of data collected over time; for example, the Gross National Product of the United States from 1945 to 2020.

treatment group. See controlled experiment.

two-sided hypothesis; two-tailed hypothesis. An alternative hypothesis asserting that the values of a parameter are different from—either greater than or less than—the value asserted in the null hypothesis. A two-sided alternative hypothesis suggests a two-sided (or two-tailed) test. See significance test; statistical hypothesis. Compare one-sided hypothesis.

two-sided test; two-tailed test. See two-sided hypothesis.

Type I error. A statistical test makes a Type I error when (1) the null hypothesis is true and (2) the test rejects the null hypothesis, i.e., there is a false positive. For example, a study of two groups may show some difference between samples from each group, even when there is no difference in the population. When a statistical test deems the difference to be significant in this situation, it makes a Type I error. See significance test; statistical hypothesis. Compare alpha; Type II error.

Type II error. A statistical test makes a Type II error when (1) the null hypothesis is false and (2) the test fails to reject the null hypothesis, i.e., there is a false negative. For example, there may not be a significant difference between samples from two groups when, in fact, the groups are different. See significance test; statistical hypothesis. Compare beta; Type I error.

unbiased estimator. An estimator that is correct on average over the possible datasets. The estimates have no systematic tendency to be high or low. Compare bias.

uniform distribution. For example, a whole number picked at random from 1 to 100 has a uniform distribution: All values are equally likely. Similarly, a uniform distribution is obtained by picking a real number at random between 0.75 and 3.25: The chance of landing in an interval is proportional to the length of the interval.

validity. Measurement validity is the extent to which an instrument measures what it is supposed to, rather than something else. The validity of a standardized test is often indicated by the correlation coefficient between the test scores and some outcome measure (the criterion variable). See content validity; differential validity; predictive validity. Compare reliability.

Study validity is the extent to which results from a study can be relied upon. Study validity has two aspects, internal and external. A study has high internal validity when its conclusions hold under the particular circumstances of the study. A study has high external validity when its results are generalizable. For example, a well-executed randomized controlled double-blind experiment performed on an unusual study population will have high internal validity because the design is good; but its external validity will be debatable because the study population is unusual.

Validity is used also in its ordinary sense: assumptions are valid when they hold true for the situation at hand.

variable. A property of units in a study, which varies from one unit to another—for example, in a study of households, household income; in a study of people, employment status (employed, unemployed, not in labor force).

variance. The square of the standard deviation. Compare standard error; covariance.

weights. See stratified random sample.

within-observer variability. Differences that occur when an observer measures the same thing twice, or measures two things that are virtually the same. Compare between-observer variability.

***z*-statistic.** A test statistic, used to perform the *z*-test. The *z*-statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested. The distinction between the *z*-statistic and the *t*-statistic is that the *z*-statistic requires that the population standard deviation be known (which is not generally the case). Some writers refer to the *t*-statistic as a *z*-statistic when the sample size is large. A *z*-statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—the estimate is too many standard errors away from its expected value. See hypothesis test; statistical hypothesis; significance test; *z*-test.

***z*-test.** A statistical test based on the *z*-statistic. The procedure is closely related to the *t*-test. The distinction between the *z*-test and the *t*-test is that the former requires that the population standard deviation is known (generally not the case). Large *z*-statistics are beyond the usual range of sampling error. For example, if z is bigger than 1.96, or smaller than –1.96, then the estimate is statistically significant at the 5% level: such values of z are hard to explain on the basis of sampling error. The scale for *z*-statistics is tied to areas under the relevant probability distribution curve.

The *z*-test arises when testing the null hypothesis that the average of a population equals a given value when the population is known to be normally distributed with a specified standard deviation. See *p*-value; significance test; statistical hypothesis.

References on Statistics and Research Methods

Nontechnical Surveys

Adam Chilton & Kyle Rozema, Trial by Numbers: A Lawyer's Guide to Statistical Evidence (2024).
David Freedman et al., Statistics (4th ed. 2007).
Darrell Huff, How to Lie with Statistics (1993).
Gregory A. Kimble, How to Use (and Misuse) Statistics (1978).
Ethan Bueno de Mesquita & Anthony Fowler, Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis (2021).
David S. Moore & William I. Notz, Statistics: Concepts and Controversies (10th ed. 2019).
Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioral Sciences (1986).
Statistics: A Guide to the Unknown (Roxy Peck et al. eds., 4th ed. 2005).
David Spiegelhalter, The Art of Statistics: Learning from Data (2019).
Hans Zeisel, Say It with Figures (6th ed. 1985).

General References

Encyclopedia of Statistical Sciences (Samuel Kotz et al. eds., 2d ed. 2005).
The Oxford Handbook of Quantitative Methods (Todd D. Little ed., 2014).
Best Practices in Quantitative Methods (Jason W. Osborne ed., 2008).

Reference Guide on Multiple Regression and Advanced Statistical Models

DANIEL L. RUBINFELD AND DAVID CARD

Daniel L. Rubinfeld, Ph.D., is Robert L. Bridges Professor of Law and Professor of Economics at the University of California, Berkeley, Emeritus and Professor of Law at New York University Law School.

David Card, Ph.D., is Class of 1950 Professor of Economics at the University of California, Berkeley.

CONTENTS

FIGURES

Introduction and Overview

This reference guide will expand on issues raised in the *Reference Guide on Statistics and Research Methods* discussing how various research designs involving the analysis of more than two variables affect the ability to draw inferences, including those regarding causation. Multiple regression techniques, which are common in litigation, will be the primary focus of this reference guide. Other less common but emerging complex statistical models, such as difference-in-differences designs, will be discussed as they relate to multiple regression and other statistical models. These research designs will be illustrated by actual and hypothetical litigation examples.

Multiple regression analysis is a statistical tool used to understand the relationship between or among two or more variables.[1] Multiple regression involves a variable to be explained—called the dependent variable or outcome variable—and additional explanatory variables that are thought to produce or be associated with changes in the dependent variable.[2] For example, a multiple regression analysis might estimate the effect of the number of years of work on salary. Salary would be the dependent variable to be explained, while the years of experience would be the explanatory variable.

Multiple regression analysis is sometimes well suited to the analysis of data when there are competing theories proposed to explain the relationship between one variable (or set of variables) and the outcome variable of interest.[3] In a case alleging gender discrimination in salaries, for example, a multiple regression analysis could be used to determine whether an average difference in salaries between women and men is attributable wholly or in part to differences between

1. A variable is anything that can take on two or more values (e.g., the daily temperature in Chicago or the salaries of workers at a factory).

2. Explanatory variables in the context of a statistical study are sometimes called independent variables or covariates. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods,* "Correlation and Regression" in this manual; *see also* section titled "What Is the Specific Question That Is Under Investigation by the Expert?" below. This reference guide also offers a brief discussion of multiple regression analysis in the section titled "More Advanced Regression Methods" below.

3. Multiple regression is one type of statistical analysis involving several variables. Other types include matching analysis, stratification, analysis of variance, probit analysis, logit analysis, discriminant analysis, and factor analysis.

the two groups in their education and experience.[4] The employer-defendant might use multiple regression to argue that salary is a function of the employee's education and experience, and the employee-plaintiff might argue that salary is also a function of the individual's sex, even taking account of education and experience. Alternatively, in an antitrust cartel damages case, the plaintiff's expert might utilize multiple regression to evaluate the extent to which the price of a product increased during the period in which the cartel was active, after accounting for costs and other variables unrelated to the cartel. The defendant's expert might use multiple regression to suggest that the plaintiff's expert has omitted several price-determining variables.

More generally, multiple regression may be useful (1) in determining whether a particular effect is present; (2) in measuring the magnitude of a particular effect; and (3) in predicting the value of the dependent variable, but for an intervening event. In a patent infringement case, for example, a multiple regression analysis could be used to estimate (1) whether the behavior of the alleged infringer affected the price of the patented product, (2) the size of the effect, and (3) what the price of the product would have been had the alleged infringement not occurred.

Over the past several decades, the use of multiple regression analysis in court has grown widely. Regression analysis has been used most frequently in cases of

4. Thus, in *Ottaviani v. State University of New York*, 875 F.2d 365, 367 (2d Cir. 1989) (citations omitted), *cert. denied*, 493 U.S. 1021 (1990), the court stated:

> In disparate treatment cases involving claims of gender discrimination, plaintiffs typically use multiple regression analysis to isolate the influence of gender on employment decisions relating to a particular job or job benefit, such as salary.

The first step in such a regression analysis is to specify all of the possible "legitimate" (i.e., nondiscriminatory) factors that are likely to significantly affect the dependent variable, and which could account for disparities in the treatment of male and female employees. By identifying those legitimate criteria that affect the decision-making process, individual plaintiffs can make predictions about what job or job benefits similarly situated employees should ideally receive, and then can measure the difference between the predicted treatment and the actual treatment of those employees. If there is a disparity between the predicted and actual outcomes for female employees, plaintiffs in a disparate treatment case can argue that the net "residual" difference represents the unlawful effect of discriminatory animus on the allocation of jobs or job benefits.

sex and race discrimination,[5] antitrust violations,[6] and cases involving class certification (under Rule 23).[7] However, there are a range of other applications,

5. Discrimination cases using multiple regression analysis are legion. *See, e.g.*, Bazemore v. Friday, 478 U.S. 385 (1986), *on remand*, 848 F.2d 476 (4th Cir. 1988); Csicseri v. Bowsher, 862 F. Supp. 547 (D.D.C. 1994) (age discrimination), *aff'd*, 67 F.3d 972 (D.C. Cir. 1995); EEOC v. Gen. Tel. Co., 885 F.2d 575 (9th Cir. 1989), *cert. denied*, 498 U.S. 950 (1990); Bridgeport Guardians, Inc. v. City of Bridgeport, 735 F. Supp. 1126 (D. Conn. 1990), *aff'd*, 933 F.2d 1140 (2d Cir.), *cert. denied*, 502 U.S. 924 (1991); Bickerstaff v. Vassar College, 196 F.3d 435, 448–49 (2d Cir. 1999) (sex discrimination); McReynolds v. Sodexho Marriott, 349 F. Supp. 2d 1 (D.D.C. 2004) (race discrimination); Hnot v. Willis Grp. Holdings Ltd., 228 F.R.D. 476 (S.D.N.Y. 2005) (gender discrimination); Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006) (sex discrimination); Coward v. ADT Sec. Sys., Inc., 140 F.3d 271, 274–75 (D.C. Cir. 1998); Smith v. Va. Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) (en banc); Hemmings v. Tidyman's Inc., 285 F.3d 1174, 1184–86 (9th Cir. 2002); Mehus v. Emporia State Univ., 222 F.R.D. 455 (D. Kan. 2004) (sex discrimination); Guiterrez v. Johnson & Johnson, 467 F. Supp. 2d 403 (D.N.J. 2006) (race discrimination); Morgan v. United Parcel Serv., 380 F.3d 459 (8th Cir. 2004) (racial discrimination); Students for Fair Admissions, Inc. v. Univ. of N.C., 567 F. Supp. 3d 580 (M.D.N.C. 2021), *cert. granted*, 142 S. Ct. 896 (2022) (race discrimination in higher education); City of Oakland v. Wells Fargo & Co., 972 F.3d 1112 (9th Cir. 2020), *vacated*, 993 F.3d 1077 (9th Cir. 2021) (racial housing discrimination); Moussouris v. Microsoft Corp., 311 F. Supp. 3d 1223 (W.D. Wash. 2018) (sex discrimination); Wal-Mart Stores, Inc. v. Dukes, 564 U.S. 338 (2011) (sex discrimination); Chen-Oster v. Goldman, Sachs & Co., 114 F. Supp. 3d 110 (S.D.N.Y. 2015) (sex discrimination); Spencer v. Va. State. Univ., 919 F.3d 199 (E.D. Va. 2019) (sex discrimination). *See also* Keith N. Hylton & Vincent D. Rougeau, *Lending Discrimination: Economic Theory, Econometric Evidence, and the Community Reinvestment Act*, 85 Geo. L.J. 237, 238 (1996) ("regression analysis is probably the best empirical tool for uncovering discrimination").

6. *E.g.*, United States v. Brown Univ., 805 F. Supp. 288 (E.D. Pa. 1992) (price fixing of college scholarships), *rev'd*, 5 F.3d 658 (3d Cir. 1993); Petruzzi's IGA Supermarkets, Inc. v. Darling-Delaware Co., 998 F.2d 1224 (3d Cir.), *cert. denied*, 510 U.S. 994 (1993); Ohio *ex rel.* Montgomery v. Louis Trauth Dairy, Inc., 925 F. Supp. 1247 (S.D. Ohio 1996); *In re* Chicken Antitrust Litig., 560 F. Supp. 963, 993 (N.D. Ga. 1980); New York v. Kraft Gen. Foods, Inc., 926 F. Supp. 321 (S.D.N.Y. 1995); Freeland v. AT&T Corp., 238 F.R.D. 130 (S.D.N.Y. 2006); *In re* Pressure Sensitive Labelstock Antitrust Litig., 566 F. Supp. 2d 363 (M.D. Pa., 2008); *In re* Linerboard Antitrust Litig., 497 F. Supp. 2d 666 (E.D. Pa. 2007) (price fixing by manufacturers of corrugated boards and boxes); *In re* Polypropylene Carpet Antitrust Litig., 93 F. Supp. 2d 1348 (N.D. Ga. 2000); *In re* OSB Antitrust Litig., No. 06-826, 2007 WL 2253418 (E.D. Pa. Aug. 3, 2007) (price fixing of Oriented Strand Board, also known as "waferboard"); *In re* TFT-LCD (Flat Panel) Antitrust Litig., 267 F.R.D. 583 (N.D. Cal. 2010); *In re* Urethane Antitrust Litig., 768 F.3d 1245 (10th Cir. 2014); *In re* Southeastern Milk Antitrust Litig., 739 F.3d 262 (6th Cir. 2014); *In re* Disposable Contact Lens Antitrust Litig., 329 F.R.D. 336 (M.D. Fla. 2018); *In re* Broiler Chicken Antitrust Litig., No. 16-C-8637, 2022 WL 1720468 (N.D. Ill. May 27, 2022).

For a broad overview of the use of regression methods in antitrust, see *ABA Antitrust*, Econometrics: Legal, Practical and Technical Issues (John Harkrider & Daniel Rubinfeld eds., 2005). *See also* Jerry Hausman et al., *Competitive Analysis with Differenciated Products*, 34 Annales D'Économie et de Statistique 159 (1994), https://doi.org/10.2307/20075951; Gregory J. Werden, *Simulating the Effects of Differentiated Products Mergers: A Practical Alternative to Structural Merger Policy*, 5 Geo. Mason L. Rev. 363 (1997). For a basic guide, *see* Michael Cragg et al., *Understanding the Econometric Tools of Antitrust—With No Math!*, 35 Antitrust Mag. (2021), https://perma.cc/P2Z2-Y9CN.

7. In *Comcast Corp. v. Behrend*, 133 S. Ct. 1426 (2013), the Court found that deficiencies in the plaintiffs' regression model precluded class certification under Rule 23(b)(3). In light of this ruling,

including census undercounts,[8] voting rights,[9] the study of the deterrent effect of the death penalty,[10] rate regulation,[11] and intellectual property.[12]

however, circuits are split as to the extent to which antitrust plaintiffs must prove that common elements predominate over individual elements. *E.g.*, *compare In re* Rail Freight Fuel Surcharge Antitrust Litig., 934 F.3d 619 (D.C. Cir. 2019) (finding the number of unharmed plaintiffs defeated predominance) *with* Olean Wholesale Grocery Coop., Inc. v. Bumble Bee Foods LLC, 31 F.4th 651 (9th Cir. 2022) (rehearing en banc) (applying a preponderance of the evidence standard to prerequisites for class certification, but finding that plaintiffs' models satisfied predomination test) *and* Kleen Prods. LLC v. Int'l Paper, 306 F.R.D. 585 (N.D. Ill. 2015) (finding that debate over the expert's multiple regression model "is a merits question that the Court does not need to resolve in order to decide whether to certify the class") *aff'd*, 831 F.3d 919 (7th Cir. 2016), *cert. denied*, 137 S. Ct. 1582 (2017). For a discussion of the use of multiple regression in evaluating class certification, see Bret M. Dickey & Daniel L. Rubinfeld, *Antitrust Class Certification: Towards an Economic Framework*, 66 N.Y.U. Ann. Surv. Am. L. 459 (2010) *and* John H. Johnson & Gregory K. Leonard, *Economics and the Rigorous Analysis of Class Certification in Antitrust Cases*, 3 J. Competition L. & Econ. 341 (2007), https://doi.org/10.1093/joclec/nhm009.

8. *See, e.g.*, City of New York v. U.S. Dep't of Commerce, 822 F. Supp. 906 (E.D.N.Y. 1993) (decision of Secretary of Commerce not to adjust the 1990 census was not arbitrary and capricious), *vacated*, 34 F.3d 1114 (2d Cir. 1994) (applying heightened scrutiny), *rev'd sub nom.*, Wisconsin v. City of New York, 517 U.S. 1 (1996); Carey v. Klutznick, 508 F. Supp. 420, 432–33 (S.D.N.Y. 1980) (use of reasonable and scientifically valid statistical survey or sampling procedures to adjust census figures for the differential undercount is constitutionally permissible), *stay granted*, 449 U.S. 1068 (1980), *rev'd on other grounds*, 653 F.2d 732 (2d Cir. 1981), *cert. denied*, 455 U.S. 999 (1982); Young v. Klutznick, 497 F. Supp. 1318, 1331 (E.D. Mich. 1980), *rev'd on other grounds*, 652 F.2d 617 (6th Cir. 1981), *cert. denied*, 455 U.S. 939 (1982).

9. Multiple regression analysis was used in suits charging that at-large area-wide voting was instituted to neutralize Black voting strength, in violation of section 2 of the Voting Rights Act, 42 U.S.C. § 1973 (1988). Multiple regression demonstrated that the race of the candidates and that of the electorate were determinants of voting. *See* Williams v. Brown, 446 U.S. 236 (1980); Rodriguez v. Pataki, 308 F. Supp. 2d 346, 414 (S.D.N.Y. 2004); United States v. Vill. of Port Chester, No. 06 Civ. 15173 (SCR), 2008 U.S. Dist. LEXIS 4914 (S.D.N.Y. Jan. 17, 2008); Meza v. Galvin, 322 F. Supp. 2d 52 (D. Mass. 2004) (violation of VRA with regard to Hispanic voters in Boston); Bone Shirt v. Hazeltine, 336 F. Supp. 2d 976 (D.S.D. 2004) (violations of VRA with regard to Native American voters in South Dakota); Georgia v. Ashcroft, 195 F. Supp. 2d 25 (D.D.C. 2002) (redistricting of Georgia's state and federal legislative districts); Benavidez v. City of Irving, 638 F. Supp. 2d 709 (N.D. Tex. 2009) (challenge of city's at-large voting scheme); Common Cause v. Rucho, 279 F. Supp. 3d 587 (M.D.N.C 2018), *rev'd*, 139 S. Ct. 2484 (2019) (challenge to partisan gerrymander); Rodriguez v. Harris County, 964 F. Supp. 2d 686 (S.D. Tex. 2013) (racially motivated gerrymandering and vote dilution claims); Luna v. County of Kern, 291 F. Supp. 3d 1088 (E.D. Cal. 2018) (racially motivated vote dilution claim). For commentary on statistical issues in voting rights cases, see, e.g., Daniel L. Rubinfeld, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991), https://doi.org/10.1177/0193841X9101500601; Stephen P. Klein et al., *Ecological Regression Versus the Secret Ballot*, 31 Jurimetrics J. 393 (1991); James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589 (1989); Arthur Lupia & Kenneth McCue, *Why the 1980s Measures of Racially Polarized Voting Are Inadequate for the 1990*s, 12 Law & Pol'y 353 (1990), https://doi.org/10.1111/j.1467-9930.1990.tb00053.x; D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want To Be?*, 47 Jurimetrics J. 115 (2007); D. James Greiner, *Re-Solidifying Racial Bloc Voting: Empirics and Legal*

Multiple regression analysis can be a source of valuable scientific testimony in litigation. When used inappropriately, however, regression analysis can confuse important issues while having little, if any, probative value. In *EEOC v. Sears, Roebuck & Co.*,[13] in which the U.S. Equal Employment Opportunity Commission (EEOC) charged Sears with discrimination against women in hiring practices, the Seventh Circuit acknowledged that "[m]ultiple regression analyses, designed to determine the effect of several independent variables on a dependent

Doctrine in the Melting Pot, 86 Ind. L.J. 447 (2011); Matt Barreto et al., *A Novel Method for Showing Racially Polarized Voting: Bayesian Improved Surname Geocoding*, 46 N.Y.U. Rev. L. & Soc. Change 1 (2022); Eric Slud et al., Ctr. for Stat. Rsch. & Methodology, U.S. Census Bureau, Statistical Methodology (2016) for Voting Rights Act, Section 203 Determinations (2018).

10. *See, e.g.*, Gregg v. Georgia, 428 U.S. 153, 184–86 (1976). For critiques of the validity of the deterrence analysis, see Nat'l Rsch. Council, Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates (Alfred Blumstein et al. eds., 1978), and updated 2012 report, Nat'l Rsch. Council, *Deterrence and the Death Penalty* 2 (Daniel S. Nagin & John V. Pepper eds., 2012) (concluding that the research is "not informative" as to any deterrent effect); Richard O. Lempert, *Desert and Deterrence: An Assessment of the Moral Bases of the Case for Capital Punishment*, 79 Mich. L. Rev. 1177 (1981); Hans Zeisel, *The Deterrent Effect of the Death Penalty: Facts v. Faith*, 1976 Sup. Ct. Rev. 317 (1976); and John Donohue & Justin Wolfers, *Uses and Abuses of Statistical Evidence in the Death Penalty Debate*, 58 Stan. L. Rev. 791 (2005).

11. *See, e.g.*, Time Warner Entertainment Co., L.P. v. FCC, 56 F.3d 151 (D.C. Cir. 1995) (challenge to FCC's application of multiple regression analysis to set cable rates), *cert. denied*, 516 U.S. 1112 (1996); Appalachian Power Co. v. EPA, 135 F.3d 791 (D.C. Cir. 1998) (challenging the EPA's application of regression analysis to set nitrous oxide emission limits); Consumers Util. Rate Advocacy Div. v. Ark. PSC, 99 Ark. App. 228 (Ark. Ct. App. 2007) (challenging an increase in non-gas rates); Qwest Corp. v. Boyle, 589 F.3d 985 (8th Cir. 2009) (challenge to the Nebraska Public Service Commission's telecoms rate setting); *In re* Rail Freight Fuel Surcharge Antitrust Litig., 725 F.3d 244 (D.C. Cir. 2013) (remanding shipping rate case, writing that the *Behrend* decision interpreted Rule 23 to "command" a hard look at regressions at the class certification stage).

12. *See* Polaroid Corp. v. Eastman Kodak Co., No. 76-1634-MA, 1990 WL 324105, at *29, *62–63 (D. Mass. Oct. 12, 1990) (damages awarded because of patent infringement), *amended by* No. 76-1634-MA, 1991 WL 4087 (D. Mass. Jan. 11, 1991); Estate of Vane v. The Fair, Inc., 849 F.2d 186, 188 (5th Cir. 1988) (lost profits were the result of copyright infringement), *cert. denied*, 488 U.S. 1008 (1989); Louis Vuitton Malletier v. Dooney & Bourke, Inc., 525 F. Supp. 2d 558, 664 (S.D.N.Y. 2007) (trademark infringement and unfair competition suit); Stone Brewing Co., LLC v. MillerCoors LLC, 445 F. Supp. 3d 1113 (S.D. Cal. 2020) (utilizing regression analysis in calculating lost profits in trademark infringement case); Navarro v. P&G, 515 F. Supp. 3d 718 (S.D. Ohio 2021) (copyright infringement case).

The use of multiple regression analysis to estimate damages has been contemplated in a wide variety of contexts. *See, e.g.*, David Baldus et al., *Improving Judicial Oversight of Jury Damages Assessments: A Proposal for the Comparative Additur/Remittitur Review of Awards for Nonpecuniary Harms and Punitive Damages*, 80 Iowa L. Rev. 1109 (1995); Talcott J. Franklin, *Calculating Damages for Loss of Parental Nurture Through Multiple Regression Analysis*, 52 Wash. & Lee L. Rev. 271 (1995); Roger D. Blair & Amanda Kay Esquibel, *Yardstick Damages in Lost Profit Cases: An Econometric Approach*, 72 Denv. U. L. Rev. 113 (1994); Daniel Rubinfeld, *Quantitative Methods in Antitrust*, *in* 1 Issues in Competition Law & Policy 723 (2008). *See also infra* note 101.

13. 839 F.2d 302 (7th Cir. 1988).

variable, which in this case is hiring, are an accepted and common method of proving disparate treatment claims."[14] However, the court affirmed the district court's findings that the "E.E.O.C.'s regression analyses did not 'accurately reflect Sears' complex, nondiscriminatory decision-making processes'" and that the "'E.E.O.C.'s statistical analyses [were] so flawed that they lack[ed] any persuasive value.'"[15] Serious questions also have been raised about the use of multiple regression analysis in census undercount cases and in death penalty cases.[16]

The Supreme Court's rulings in *Daubert* and *Kumho Tire* have encouraged parties to raise questions about the admissibility of multiple regression analyses.[17] Because multiple regression is a well-accepted scientific methodology, courts have frequently admitted testimony based on multiple regression studies, in some cases

14. *Id.* at 324 n.22.

15. *Id.* at 348, 351 (quoting EEOC v. Sears, Roebuck & Co., 628 F. Supp. 1264, 1342, 1352 (N.D. Ill. 1986)). The district court commented specifically on the "severe limits of regression analysis in evaluating complex decision-making processes." 628 F. Supp. at 1350.

16. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Correlation and Regression" and see sections titled "Choosing the Additional Explanatory Variables" and "Choosing the Dependent Variable" below.

17. Daubert v. Merrill Dow Pharms., Inc. 509 U.S. 579 (1993); Kumho Tire Co. v. Carmichael, 526 U.S. 137, 147 (1999) (expanding the *Daubert* application to nonscientific expert testimony). For example, analysis conducted by PricewaterhouseCoopers on challenges to financial expert witnesses found an approximately eight-fold increase since 2000. PricewaterhouseCoopers, *Daubert* Challenges to Financial Experts: A Yearly Study of Trends and Outcomes 4 (2000–2021), https://perma.cc/38B5-GTTB. There was a 19% increase of challenges between 2020 and 2021. *Id.* Of those, eighty-nine challenges (33%) resulted in partial or full exclusion of the expert. *Id.*

Circuit courts have developed extensive case law on the issue of applying the *Daubert* standard to expert testimony introduced before the class-action certification stage. Dictum in *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338 (2011), suggested that the district court erred in not applying *Daubert* to expert testimony, and while the Court certified the question in *Comcast Corp. v. Behrend*, 133 S. Ct. 1426 (2013), it did not reach the merits to decide it. In the resulting vacuum, a plurality of circuits held that district courts must submit expert testimony to *Daubert* scrutiny. *See* Prantil v. Arkema Inc., 986 F.3d 570, 575–76 (5th Cir. 2021) ("if an expert's opinion would not be admissible at trial, it should not pave the way for certifying a proposed class"); citing *In re* Blood Reagents Antitrust Litig., 783 F.3d 183, 187 (3d Cir. 2015) ("We join certain of our sister courts to hold that a plaintiff cannot rely on challenged expert testimony, when critical to class certification, to demonstrate conformity with Rule 23 unless the plaintiff also demonstrates, and the trial court finds, that the expert testimony satisfies the standard set out in *Daubert*"); Sher v. Raytheon Co., 419 F. App'x 887, 890–91 (11th Cir. 2011) ("Here the district court refused to conduct a *Daubertlike* critique of the proffered experts's qualifications. This was error."); Am. Honda Motor Co. v. Allen, 600 F.3d 813, 815–16 (7th Cir. 2010) ("We hold that when an expert's report or testimony is critical to class certification, as it is here, . . . a district court must conclusively rule on any challenge to the expert's qualifications or submissions prior to ruling on a class certification motion."); Grodzitsky v. Am. Honda Motor Co., 957 F.3d 979, 984 (9th Cir. 2020) ("[I]n evaluating challenged expert testimony in support of class certification, a district court should evaluate admissibility under the standard set forth in *Daubert*.").

over the strong objection of one of the parties.[18] On some occasions courts have excluded expert testimony because of a failure to utilize a multiple regression methodology.[19] On other occasions, courts have rejected regression studies that did not have an adequate foundation or research design with respect to the issues at hand.[20]

In interpreting the results of a multiple regression analysis, it is important to distinguish between correlation and causality. Two variables are correlated—that is, associated with each other—when the events associated with the variables occur more frequently together than one would expect by chance. For example, if higher salaries are associated with a greater number of years of work experience, and lower salaries are associated with fewer years of experience, there is a positive correlation between salary and number of years of work experience. However, if higher salaries are associated with less experience, and lower salaries are associated with more experience, there is a negative correlation between the two variables.

A correlation between two variables does not necessarily imply that one event causes the second. One common explanation for situations where two variables are correlated but there is no causal connection is spurious correlation.[21] Spurious correlation arises when two variables move together because they are both caused by a third, unexamined variable. For example, there might be a negative correlation between the age of certain skilled employees of a computer company and their salaries. One should not conclude from this correlation that the employer has necessarily discriminated against the employees based on their age. A third, unexamined variable, such as the level of the employees' technological skills, could explain differences in productivity and, consequently, differences in salary.[22] Or

18. *See* Newport Ltd. v. Sears, Roebuck & Co., Civ. Action No. 86-2319 Section "K," 1995 U.S. Dist. LEXIS 7652 (E.D. La. May 26, 1995). *See also* Petruzzi's IGA Supermarkets, Inc. v. Darling-Delaware Co., 998 F.2d 1224, 1240–47 (3d Cir.), *cert. denied*, 510 U.S. 994 (1993) (finding that the district court abused its discretion in excluding multiple regression-based testimony and reversing the grant of summary judgment to two defendants).

19. *See, e.g.*, *In re* Exec. Telecard Ltd. Sec. Litig., 979 F. Supp. 1021 (S.D.N.Y. 1997); *but see* United States v. Valencia, 600 F.3d 389 (5th Cir. 2010) (ruling that the lack of a regression analysis was a matter of weight, not admissibility, of expert testimony).

20. See *City of Tuscaloosa v. Harcros Chems., Inc.*, 158 F.3d 548 (11th Cir. 1998), in which the court ruled plaintiffs' regression-based expert testimony inadmissible and granted summary judgment to the defendants. See also *Am. Booksellers Ass'n v. Barnes & Noble, Inc.*, 135 F. Supp. 2d 1031, 1041 (N.D. Cal. 2001), in which a model was said to contain "too many assumptions and simplifications that are not supported by real-world evidence"; *see also* Obrey v. Johnson, 400 F.3d 691 (9th Cir. 2005).

21. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "What Inferences Can Be Drawn from the Data?" section, in this manual.

22. *See, e.g.*, Sheehan v. Daily Racing Form Inc., 104 F.3d 940, 942 (7th Cir.) (rejecting plaintiff's age discrimination claim because statistical study showing correlation between age and retention ignored the "more than remote possibility that age was correlated with a legitimate job-related qualification"), *cert. denied*, 521 U.S. 1104 (1997).

consider a patent infringement case in which increased sales of an allegedly infringing product are associated with a lower price of the patented product.[23] This correlation would be spurious if the two products have their own noncompetitive market niches and the lower price is the result of a decline in the production costs of the patented product.

Raising the possibility of a spurious correlation will typically not be enough to dispose of a statistical argument asserting causality. It will normally be necessary to show that the third factor that is alleged to cause the spurious correlation is itself relevant. For example, a statistical showing of a relationship between technological skills and worker productivity might be required in the age discrimination example above.[24]

In most litigation settings involving statistical analysis, causal relationships are specified within the framework of an underlying causal theory that explains the relationship between the two variables. Even when an appropriate theory has been identified, causality cannot be inferred from the theory alone. One must also look for empirical evidence that a causal relationship exists. Conversely, the fact that two variables are correlated does not guarantee the existence of the causal relationship posited in each theory; it could be that the regression model—an approximation of the underlying causal theory—does not reflect the correct interplay among the explanatory variables. Likewise, the absence of correlation does not guarantee that a causal relationship does not exist. Lack of correlation can occur, even when there is a true causal effect from one variable to another if (1) there are insufficient data, (2) the data are measured inaccurately, (3) the data do not allow multiple causal relationships to be sorted out, or (4) the model is specified wrongly because of the omission of a variable or variables that are related to the variable of interest.

In recent years, new methodologies have broadened our ability to draw causal inferences from a variety of data sources. This reference guide includes an explanation of why the randomized controlled trial (RCT) method is widely accepted in the scientific community as offering the strongest evidence of causal relationships. This reference guide also includes a discussion of a standard framework that delineates the precise meaning of "causality" in an RCT, and the extension

23. In some cases, there are statistical tests that allow one to reject claims of causality. For a brief description of these tests, which were developed by Jerry Hausman, see Robert S. Pindyck & Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts* § 7.5 (4th ed. 1997).

24. *See, e.g.*, Allen v. Seidman, 881 F.2d 375 (7th Cir. 1989) (judicial skepticism was raised when the defendant did not submit a logistic regression incorporating an omitted variable; defendant's attack on statistical comparisons must also include an analysis that demonstrates that the comparisons are flawed). The appropriate requirements for the defendant's showing of spurious correlation could, in general, depend on the discovery process. *See, e.g.*, Boykin v. Georgia Pac. Co., 706 F.2d 1384 (1983) (criticism of a plaintiff's analysis for not including omitted factors, when plaintiff considered all information on an application form, was inadequate).

of that framework to other nonexperimental research designs.[25] In the case of an RCT, the random assignment of individuals to different subgroups that receive different "treatments" helps ensure that the subgroups are composed of similar individuals.[26] In that case, differences in the outcome of interest between the subgroups can be attributed to differences in the assigned treatments—since presumably that is the only factor that varies across otherwise randomly determined groups. In the absence of random assignment, individuals with different characteristics typically choose their own groups (or are assigned to different groups by some other party or process). This self-selection into different groups makes it difficult to determine if the differences across groups are caused by differences in exposure to the explanatory variable of interest or to other differences between the groups.

Occasionally, naturally occurring experiments will arise, in which the ordinary operation of a system results in a close approximation of random assignment in a controlled trial. For example, access to an innovative postconviction job training program may be allocated based on a lottery, thereby approximating the random assignment process of a controlled trial. Comparing employment records of those assigned by lottery to the job training program with records of those not assigned by lottery to the program should allow an assessment of the impact of the job training program on subsequent employment history. Such an assessment will be more accurate than simple comparisons between program participants and nonparticipants in settings where participants are self-selected, since often those who self-select into a program are more highly motivated or otherwise different from those who do not. In such cases it is difficult to sort out the causal effects of the program from differences attributable to the motivation (and other characteristics) of the individuals who selected themselves into the two groups.

In the absence of randomized controlled trials or naturally occurring experiments, complex statistical models utilizing and building on regression methods have been developed to allow for drawing causal inferences. Such models allow individuals to sort themselves into different groups and use various forms of statistical analysis as a means of adjusting for such naturally occurring differences, while still allowing for meaningful assessment of differences across the groups. This reference guide will explore how such statistical analyses and adjustments attempt to ensure that the differences identified by the analyses are because of

25. Randomized *clinical* trials are leading examples of RCTs. Many RCTs, however, are not conducted in a clinical environment. See, e.g., the summary of randomized social experiments in David H. Greenberg & Mark Shroder, *Digest of Social Experiments* (3d ed. 2004).

26. In the experimental literature, the different treatment groups are sometimes referred to as "treatment arms." In the simplest experiment, one treatment arm receives no treatment (or a placebo treatment), while the other arm receives the treatment of interest, such as a new medicine or a new welfare benefit program.

exposure to different research circumstances and not because of differences arising from the self-classification of individuals.

There is a tension between any attempt to reach conclusions with near certainty and the inherently probabilistic nature of multiple regression analysis. In general, multiple regression allows for the expression of uncertainty in terms of probabilities. The reality that statistical analysis generates probabilities concerning relationships rather than certainty should not be seen as an argument against the use of statistical evidence or, worse, as a reason not to admit that there is uncertainty at all. The only alternative might be to use less reliable anecdotal evidence. Instead, the probabilistic nature of the discipline allows for the expression of different levels of confidence in a set of outcomes, upon which a trier of fact may base a decision.

This reference guide addresses several procedural and methodological issues that are relevant in considering the admissibility of, and the weight to be accorded to, the findings of multiple regression and related advanced statistical analyses. It also suggests some standards of reporting and analysis that an expert presenting analyses might be expected to meet.

A brief overview of the sections of this guide: "Research Design: Model Specification" discusses research design—how the basic regression framework can be used to sort out alternative theories about a case. The guide discusses the importance of choosing the appropriate specification of the regression and regression-related models and raises the issue of whether these methods are appropriate for the case at issue. "Interpreting Multiple Regression Results" accepts the regression framework and concentrates on the interpretation of the multiple regression results from both a statistical and a practical point of view. It emphasizes the distinction between regression results that are statistically significant and results that are meaningful to the trier of fact (i.e., practically significant). It also points to the importance of evaluating the robustness of regression analyses, that is, seeing the extent to which the results are sensitive to changes in the underlying assumptions of the regression model. "Causal Analysis and Research Designs" describes a variety of regression-related methodologies that serve as the foundation for causal analysis. Causal analysis methods allow experts to make inferences about but-for worlds—hypothetical worlds that would exist absent alleged wrongful behavior. "More Advanced Regression Methods" offers a brief overview of fixed-effects and random-effects regression models as well as a discussion of model selection methods.

"The Expert" briefly discusses the qualifications of experts and suggests a potentially useful role for court-appointed neutral experts. "Presentation of Statistical Evidence" emphasizes procedural aspects associated with use of the data underlying regression analyses. It encourages greater pretrial efforts by the parties to attempt to resolve disputes over statistical studies.

Throughout the main body of this reference guide, hypothetical examples are used as illustrations. Moreover, the basic mathematics of multiple regression has been kept to a bare minimum. To achieve that goal, the more formal

description of the multiple regression framework has been placed in the Appendix. The Appendix is self-contained and can be read before or after the text. The Appendix also includes further details with respect to the examples used in the body of this reference guide.

Research Design: Model Specification

Multiple regression allows the testifying expert to choose among alternative theories or hypotheses and assists the expert in distinguishing correlations between variables that are plainly spurious from those that may reflect valid relationships.

What Is the Specific Question That Is Under Investigation by the Expert?

Research begins with a clear formulation of a research question. The data to be collected and analyzed must relate directly to this question; otherwise, appropriate inferences cannot be drawn from the statistical analysis. For example, if the question at issue in a patent infringement case is what price the plaintiff's product would have been but for the sale of the defendant's infringing product, sufficient data must be available to allow the expert to account statistically for important factors that determine the price of the product.

What Model Should Be Used to Evaluate the Question at Issue?

Model specification involves several steps, each of which is fundamental to the success of the research effort. Ideally, a multiple regression analysis builds on a theory that describes the variables to be included in the study. A typical regression model will include one or more dependent variables, each of which is believed to be causally related to a series of explanatory variables. Because we cannot be certain that the explanatory variables are themselves unaffected or independent of the influence of the dependent variable (at least at the point of initial study), the explanatory variables are often termed *covariates*. Covariates are known to have an association with the dependent or outcome variable, but causality remains an open question.

For example, the theory of labor markets might lead one to expect that salaries in an industry are related to workers' education, experience, and training. A belief that there is gender discrimination in setting salaries would lead one to create a model in which the dependent variable is a measure of workers' salaries,

and the list of covariates includes an indicator for female gender in addition to measures of training, experience, and education.

We can imagine an alternative world in which an analysis of discrimination in pay setting (or any other issue) might be accomplished through a "natural experiment," in which for some reason a group of male and female workers who were known to be equally productive were randomly assigned to a variety of employers in an industry under study and asked to fill positions requiring identical experience and skills. In this design, where any difference in salaries could only be a result of discrimination, it would be possible to draw clear and direct inferences from an analysis of salary data. Unfortunately, the opportunity to analyze natural experiments or conduct randomized controlled trials is rarely available to experts in the context of legal proceedings. In the real world, experts must do their best to interpret the results of the available real-world data, recognizing that it is often impossible to control all factors that might affect worker salaries or other outcomes of interest.[27]

Models are often characterized in terms of parameters (numerical characteristics of the model). In the labor-market discrimination example, one parameter might reflect the increase in salaries associated with each additional year of prior job experience. Another parameter might reflect the difference in salaries associated with jobs in urban versus nonurban areas. Multiple regression uses a sample, or a selection of data, from the population (all the units of interest) to obtain estimates of the values of the parameters of the model. An estimate associated with a particular explanatory variable is an estimated regression coefficient.

Failure to develop the proper theory, failure to choose the appropriate variables, or failure to choose the correct form of the model can substantially bias the statistical results—that is, create a systematic tendency for an estimate of a model parameter to be too high or too low.

Choosing the Dependent Variable

The variable to be explained, the dependent variable, should be the appropriate variable for analyzing the question at issue.[28] Suppose, for example, that pay

27. In the literature on natural and quasi-experiments, the explanatory variables are characterized as "treatments" and the dependent variable as the "outcome." For a review of natural experiments in the criminal justice arena, see David P. Farrington, *A Short History of Randomized Experiments in Criminology*, 27 Evaluation Rev. 218–27 (2003), https://doi.org/10.1177/0193841X03027003002.

28. In multiple regression analysis, the dependent variable is often a continuous variable that takes on a range of numerical values (like a person's salary or a test score). When the dependent variable is categorical, taking on only two or three values, modified forms of multiple regression, such as probit analysis or logit analysis, are appropriate. For an example of the use of the latter, see *EEOC v. Sears, Roebuck & Co.*, 839 F.2d 302, 325 (7th Cir. 1988) (*EEOC* used logit analysis to

discrimination among hourly workers is a concern. One choice for the dependent variable is the hourly wage rate of the employees, while another choice is the annual salary. The distinction is important, because annual salary differences may in part result from differences in hours worked. If the number of hours worked is the product of worker preferences and not discrimination, the hourly wage is a good choice. If the number of hours worked is related to the alleged discrimination, annual salary is the more appropriate dependent variable to choose.[29]

Choosing the Explanatory Variable That Is Relevant to the Question at Issue

The explanatory variable that allows the evaluation of alternative hypotheses must be chosen appropriately. Thus, in a discrimination case, the variable of interest may be the race or sex of the individual. In an antitrust case, it may be a variable that takes on the value 1 to reflect the presence of the alleged anticompetitive behavior and the value 0 otherwise.[30]

Choosing the Additional Explanatory Variables

An attempt should be made to identify additional known or hypothesized explanatory variables, some of which are measurable and may support alternative substantive hypotheses that can be accounted for by the regression analysis. For example, in a discrimination case, a measure of the skills of the workers may provide an alternative explanation—lower salaries may have been the result of inadequate skills.[31]

measure the impact of variables such as age, education, job-type experience, and product-line experience on the female percentage of commission hires).

29. In job systems in which annual salaries are tied to grade or step levels, the annual salary corresponding to the job position could be the more appropriate dependent variable.

30. Explanatory variables may vary by type, which will affect the interpretation of the regression results. Thus, some variables may be continuous, and others may be categorical.

31. In *James v. Stockham Valves*, 559 F.2d 310 (5th Cir. 1977), the Court of Appeals rejected the employer's claim that skill level rather than race determined assignment and wage levels, noting the circularity of defendant's argument. In *Ottaviani v. State University of New York*, 679 F. Supp. 288, 306–08 (S.D.N.Y. 1988), *aff'd*, 875 F.2d 365 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the court ruled (at the liability phase of the trial) that the university showed that there was no discrimination in either placement into initial rank or promotions between ranks, and so rank was a proper variable in multiple regression analysis to determine whether women faculty members were treated differently than men. *Cf.* Bennett v. Nucor Corp., 656 F.3d 802, 817–18 (8th Cir. 2011) (faulting plaintiff's expert for omitting experience and qualification variables).

It is neither realistic nor useful to include all possible variables that might influence the dependent variable in a regression model; some cannot be measured, and others may make little difference.[32] If a preliminary analysis shows the unexplained portion of the multiple regression to be unacceptably high, the expert may seek to discover whether some previously undetected variable is missing from the analysis.[33]

Failure to include a major explanatory variable that is correlated with the explanatory variable of interest in a regression model is an especially serious concern. Because the parameters of a regression model are selected to maximize the

However, in *Trout v. Garrett*, 780 F. Supp. 1396, 1414 (D.D.C. 1991), the court ruled (in the damage phase of the trial) that the extent of civilian employees' prehire work experience was not an appropriate variable in a regression analysis to compute back pay in employment discrimination. According to the court, including the prehire level would have resulted in a finding of no sex discrimination, despite a contrary conclusion in the liability phase of the action. *Id. See also* Stuart v. Roache, 951 F.2d 446 (1st Cir. 1991) (allowing only three years of seniority to be considered as the result of prior discrimination), *cert. denied*, 504 U.S. 913 (1992). Whether a particular variable reflects "legitimate" considerations or itself reflects or incorporates illegitimate biases is a recurring theme in discrimination cases. *See, e.g.*, Moussouris v. Microsoft Corp., 311 F. Supp. 3d 1223, 1238 (W.D. Wash. 2018) (finding as appropriate the exclusion of two variables potentially "tainted" by gender bias). *See also* Smith v. Va. Commonwealth Univ., 84 F.3d 672, 677 (4th Cir. 1996) (en banc) (suggesting that whether "performance factors" should have been included in a regression analysis was a question of material fact); *id.* at 681–82 (Luttig, J., concurring in part) (suggesting that the failure of the regression analysis to include "performance factors" rendered it so incomplete as to be inadmissible); *id.* at 690–91 (Michael, J., dissenting) (suggesting that the regression analysis properly excluded "performance factors"); *see also* Diehl v. Xerox Corp., 933 F. Supp. 1157, 1168 (W.D.N.Y. 1996). Other times, the inclusion or exclusion of performance factors as a potentially explanatory variable is a question for the trier of fact. *See, e.g.*, Chi. Teachers Union, Local 1 v. Bd. of Educ. of Chi., No. 12-C-10311, 2020 U.S. Dist. LEXIS 32351, at *19–20 (N.D. Ill. Feb. 25, 2020) ("It is for the trier of fact to determine whether [the expert's] failure to account for academic performance in his regressions renders them less probative."). *See also* Crawford v. Newport News Indus. Corp., No. 4:14-cv-130, 2017 U.S. Dist. LEXIS 118879 (E.D. Va. July 28, 2017) (excluding expert's report that lacked control variable for "specific job classification"); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248 (4th Cir. 2005) (affirming district court's decision to exclude regression analysis that failed to compare similarly situated workers). A challenge that a model omitted certain "non-discriminatory" variables must be able to find differences themselves to sustain the challenge. *See* Buchanan v. Tata Consultancy Servs., No. 15-cv-01696-YGR, 2017 U.S. Dist. LEXIS 212170 (N.D. Cal. Dec. 27, 2017).

32. The summary effect of the excluded variables shows up as a random-error term in the regression model, as does any modeling error. See Appendix, below, for details. *But see* David W. Peterson, *Reference Guide on Multiple Regression*, 36 Jurimetrics J. 213, 214 n.2 (1996) (review essay) (asserting that "the presumption that the combined effect of the explanatory variables omitted from the model are uncorrelated with the included explanatory variables" is "a knife-edge condition . . . not likely to occur").

33. A very low *R*-squared (R^2) is one indication of an unexplained portion of the multiple regression model that is unacceptably high. However, the inference that one makes from a particular value of R^2 will depend, of necessity, on the context of the issues and particular datasets that are under study. For reasons discussed in the Appendix, a low R^2 does not necessarily imply a poor model (and vice versa).

ability of the model to predict the outcome variable using only the included variables, such an omission may cause an included variable to be incorrectly credited with an effect caused by the excluded variable.[34] Such a situation is called omitted variables bias. In this context, bias is a statistical term referring to the difference between the likely (or expected) parameter value that will be estimated in the (flawed) regression model and the true causal effect of the explanatory variable of interest on the outcome variable. A valid regression model will yield unbiased parameter values.

In general, the existence of omitted variables that are likely to be correlated with both the dependent variable and the key explanatory variable of interest in the analysis (such as gender or race in a discrimination analysis) reduce the probative value of the regression analysis. The importance of omitting a relevant variable depends on the strength of the relationship between the omitted variable and the dependent variable and the strength of the correlation between the omitted variable and the explanatory variable of interest. Other things being equal, the greater the correlation between the omitted variable and the variable of interest, the greater the bias caused by the omission. As a result, the omission of an important variable may lead to inferences made from a regression analysis that are incorrect or do not assist the trier of fact.[35]

34. Technically, the omission of explanatory variables that are correlated with the variable of interest can cause biased estimates of regression parameters.

35. *See* Bazemore v. Friday, 751 F.2d 662, 671–72 (4th Cir. 1984) (upholding the district court's refusal to accept a multiple regression analysis as proof of discrimination by a preponderance of the evidence, the court of appeals stated that, although the regression used four variable factors (race, education, tenure, and job title), the failure to use other factors, including pay increases that varied by county, precluded their introduction into evidence), *aff'd in part*, *vacated in part*, 478 U.S. 385 (1986).

Note, however, that in *Sobel v. Yeshiva University*, 839 F.2d 18, 33, 34 (2d Cir. 1988), *cert. denied*, 490 U.S. 1105 (1989), the court made clear that "a [Title VII] defendant challenging the validity of a multiple regression analysis [has] to make a showing that the factors it contends ought to have been included would weaken the showing of salary disparity made by the analysis" by making a specific attack and "a showing of relevance for each particular variable it contends . . . ought to [be] includ[ed]" in the analysis, rather than by simply attacking the results of the plaintiffs' proof as inadequate for lack of a given variable. *See also* Smith v. Va. Commonwealth Univ., 84 F.3d 672 (4th Cir. 1996) (en banc) (finding that whether certain variables should have been included in a regression analysis is a question of fact that precludes summary judgment); Freeland v. AT&T, 238 F.R.D. 130, 145 (S.D.N.Y. 2006) ("[o]rdinarily, the failure to include a variable in a regression analysis will affect the probative value of the analysis and not its admissibility").

Also, in *Bazemore v. Friday*, the Court, declaring that the Fourth Circuit's view of the evidentiary value of the regression analyses was plainly incorrect, stated that "[n]ormally, failure to include variables will affect the analysis' probativeness, not its admissibility. Importantly, a regression analysis that includes less than all measurable variables may serve to prove a plaintiff's case." 478 U.S. 385, 400 (1986) (footnote omitted). Circuits continue to follow this evidentiary ruling. *See* Kurtz v. Costco Wholesale Corp., 818 Fed. App'x 57 (2d Cir. 2020) *and* Karlo v. Pittsburgh Glass Works, LLC, 849 F.3d 61 (3d Cir. 2017); *but see In re* Scrap Metal Antitrust Litig., 527 F.3d

Omitted variables that are not correlated with the variable of interest are, in general, less of a concern, because the parameter that measures the effect of the variable of interest on the dependent variable will be estimated without bias. Suppose, for example, that the effect of a policy introduced by the courts to encourage spouses to pay child support has been tested by randomly choosing some cases to be handled according to current court policies and other cases to be handled according to a new, more stringent policy. The effect of the new policy might be measured in a multiple regression using payment success as the dependent variable and a variable indicating whether the old or new policy was in effect as the explanatory variable (1 if the new program was assigned; 0 if it was not). Failure to include an explanatory variable that reflected the age of the husbands involved in the program would not affect the court's evaluation of the new policy, because men of any given age are as likely to be affected by the old policy as they are the new policy. Randomly applying the two policies to each case has ensured that the omitted age variable is not correlated with the policy variable.

Bias caused by the omission of an important variable that is related to the included variables of interest can be a serious problem.[36] Nonetheless, it is possible for the expert to account for bias qualitatively if the expert has knowledge (even if not quantifiable) about the relationship between the omitted variable and the explanatory variable. Suppose, for example, that the plaintiff's expert in a sex discrimination pay case is unable to obtain quantifiable data that reflect the skills necessary for a job, but it is known that, on average, women are more skillful than men. Suppose also that a regression analysis of the wage rate of employees (the dependent variable) on years of experience and a variable reflecting the sex of each employee (the key explanatory variable) suggests that men are paid substantially more than women with the same experience. Because differences in skill levels have not been accounted for, the expert may reasonably conclude that the wage difference measured by the regression is a conservative estimate of the true discriminatory effect on wages.[37]

The precision of the measure of the effect of a variable of interest on the dependent variable is also important.[38] In general, the precision of the estimated

517, 530 (6th Cir. 2008) ("That is not to say that a significant error in application will never go to the admissibility, as opposed to the weight, of the evidence.").

36. *See also* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "What Inferences Can Be Drawn from the Data?" section, in this manual.

37. The inclusion of potentially cherry-picked skill data can likewise serve as a red flag for courts. In *Moussouris v. Microsoft Corp.*, 311 F. Supp. 3d 1223, 1246 (W.D. Wash. 2018), the court found that by omitting younger, less experienced members of the relevant employee cohort, defendant's expert relied on "unrepresentative and thus insufficient" data and excluded her testimony under Federal Rule of Evidence 702.

38. A more precise estimate of a parameter is an estimate with a smaller standard error. The confidence interval associated with a more precise estimate will be smaller. See Appendix, below, for details.

coefficient in a multiple regression model depends on three factors: (1) the sample size (larger samples give more precise results); (2) the extent to which the model successfully explains the dependent variable (higher explanatory power gives more precise results); and (3) the extent to which the variable of interest varies independently of the other covariates in the model (more independence gives more precise results). Sometimes the inclusion of additional covariates can improve the explanatory power of the model and raise the precision of the estimated coefficient of the variable of interest. Including explanatory variables that are irrelevant (i.e., that do not help increase the explanatory power of the model) will reduce the precision of the estimated coefficients. This can cause concern when the sample size is small, but it is not likely to be of great consequence when the sample size is large.

Choosing the Form of the Multiple Regression Model

Choosing the proper set of variables for a multiple regression model does not complete the modeling exercise. The expert must also choose the proper form of the regression model. The most frequently selected form is the linear regression model allowing separate coefficients for each of the explanatory variables in the model (described in the Appendix). In such a model, the magnitude of the change in the dependent variable that is associated with the change in any of the explanatory variables is the same no matter what the level of the explanatory variables. For example, one additional year of experience might add $5,000 to salary, regardless of the employee's sex or their previous experience.

In some instances, however, there may be reason to believe that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change. In these instances, the expert should consider the use of a more flexible model. Suppose, for example, that having two bathrooms in a house adds twice as much value as having only one bathroom. However, the value of a third bathroom is substantially lower than the value of the first or second bathroom, and even more so for the fourth bathroom. This might be accounted for by having the price of a house be a function of (a) the number of bathrooms and (b) the square of the number of bathrooms. We would expect that the first variable would have a positive effect on price, whereas the second variable would have a negative effect. Failure to account for relationships such as this one can lead to either overstatement or understatement of the effect of a change in the value of an explanatory variable on the value of a dependent variable.

Another source of flexibility is to include interactions among the individual explanatory variables. An interaction variable is the product of two other variables that are included in the multiple regression model. The interaction variable allows the expert to consider the possibility that the effect of a change in one variable on the dependent variable may change as the level of another

explanatory variable changes. For example, in a salary discrimination case, a variable of interest might be the sex of the individual, allowing for the possibility that there are wage differences between men and women. However, a more complete description might also allow for the inclusion of a term that interacts (multiplies) a variable measuring experience with the variable representing the sex of the employee (1 if a female employee; 0 if a male employee). This allows the expert to test whether the sex differential varies with the level of experience. A significant negative estimate of the parameter associated with the sex variable would suggest that inexperienced women are discriminated against, whereas a significant negative estimate of the interaction parameter suggests that the extent of discrimination increases with experience.[39]

There may be cases where a model includes *both* the independent variable of interest and its interaction with some other variable—as in the previous example where female sex is included as one variable and female sex interacted with experience is included as a second variable. Then, one must account for the parameters for both terms to infer the effect of the variable of interest on any group.

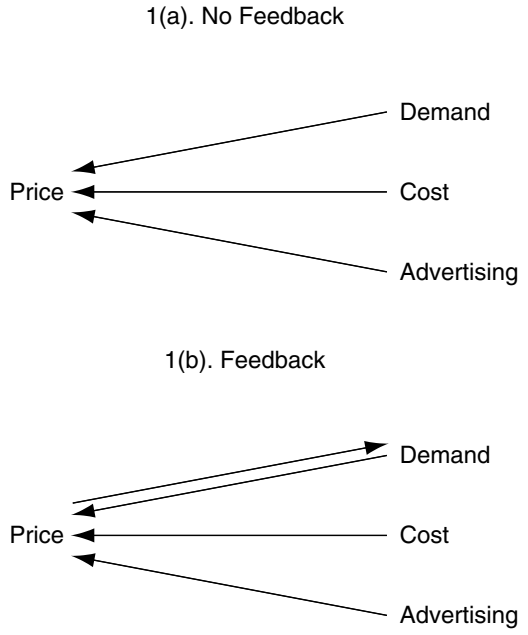
Dealing with Endogeneity

In a multiple regression framework, the expert often assumes that changes in an explanatory variable affect the dependent variable but that changes in the dependent variable do not affect the explanatory variable—that is, there is no feedback.[40] In cases where there is no feedback, and no spurious correlation owing to unobserved factors that influence both the explanatory and outcome variables, all of the correlation between the explanatory variable and the dependent outcome variable arises from the effect of the former on the latter, and not vice versa. Under

39. For further details concerning interactions, see the Appendix, below. Note that in *Ottaviani v. State Univ. of N.Y.*, 875 F.2d 365, 367 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the defendant relied on a regression model in which a dummy variable reflecting gender appeared as an explanatory variable. The female plaintiff, however, used an alternative approach in which a regression model was developed for men only (the alleged protected group). The salaries of women predicted by this equation were then compared with the actual salaries; a positive difference would, according to the plaintiff, provide evidence of discrimination. For an evaluation of the methodological advantages and disadvantages of this approach, see Joseph L. Gastwirth, *A Clarification of Some Statistical Issues in* Watson v. Fort Worth Bank & Trust, 29 Jurimetrics J. 267 (1989).

40. The "no feedback" assumption is not sufficient to ensure that the estimated coefficient on the variable of interest will be unbiased. It must be assumed further that there is no correlation between any omitted variables in the regression equation (as reflected in the error term) and any omitted variables in an equation in which the variable of interest and the outcome variable are reversed (i.e., no spurious correlation). The "no feedback" assumption is especially important in litigation because it is possible for the defendant (if responsible, for example, for price fixing or discrimination) to affect the values of the explanatory variables and thus to bias the usual statistical tests that are used in multiple regression.

Figure 1. Feedback.



these assumptions, the estimated coefficient of the explanatory variable in the multiple regression model represents the causal effect of that variable on the outcome variable. If the "no feedback" assumption is false, or there is a spurious correlation from unobserved factors that affect both variables, the estimated effect of the explanatory variable on the outcome variable is likely to be biased, causing the expert and the trier of fact to reach the wrong conclusion about the true magnitude of the causal effect.

In many situations—particularly those involving the modeling of prices and quantities of a product sold in a market—there is two-way feedback between the outcome variable and the explanatory variable of interest. As a result, if the expert does not take this more complex relationship into account, the regression coefficient on the variable of interest could be either too high or too low. When such two-way feedback is present (or there is a spurious correlation attributable to unobserved factors that affect both variables) the outcome variable and the explanatory variable are said to be "simultaneously determined" (i.e., their relationship

is characterized as one of simultaneity), and the covariate that is believed to be affected by simultaneity is said to be endogenous.[41]

Figure 1 illustrates this point. In Figure 1(a), the dependent variable, Price, is explained through a multiple regression framework by three covariate explanatory variables—demand, cost, and advertising—with no feedback. Each of the three covariates is assumed to affect price causally, while price is assumed to have no effect on the three covariates. However, in Figure 1(b), there is feedback, because price affects demand and demand, cost, and advertising affect price. Cost and advertising, however, are not affected by price. In this case both price and demand are jointly determined endogenous variables, that is, each has a causal effect on the other.

As a rule, there are no direct statistical tests for determining the direction of causality; rather, the expert, when asked, should be prepared to defend their assumption based on an understanding of the underlying behavior evidence relating to the businesses or individuals involved.[42]

Although there is no single approach that is entirely suitable for estimating models when the dependent variable affects one or more of the explanatory variables, one possibility is to drop the questionable variable from the regression to determine whether the variable's exclusion makes a difference. If it does not, the issue becomes moot. Another approach is to expand the multiple regression model by adding one or more equations that explain the relationship between the variable of interest (the explanatory variable) and an outcome variable (the dependent variable). A third approach is to find an instrumental variable—a variable that substantially affects the variable of interest, does not directly affect the outcome variable, and is uncorrelated with any omitted explanatory variables that might affect the dependent variable.

Suppose, for example, that in a salary-based racial discrimination suit the defendant's expert considers employer-evaluated test scores to be an appropriate explanatory variable for the dependent variable, the wage rate. If the plaintiff were to provide information that the employer adjusted the test scores in a manner that penalized the wages of African-American workers, the assumption that wages were determined by test scores alone might be invalid. It might be a totally inappropriate covariate, or it might be endogenous. If the test-score variable is inappropriate, it should be removed from consideration. If endogenous,

41. For discussions of endogeneity problems, see *Conrad v. Jimmy John's Franchise, LLC*, No. 18-CV-00133, 2021 U.S. Dist. LEXIS 84039 (S.D. Ill. Apr. 26, 2021); *In re Nat'l Prescription Opiate Litig.*, No. 1:17-MD-2804, 2019 U.S. Dist. LEXIS 141129 (N.D. Ohio Aug. 20, 2019); and *In re High-Tech Employ. Antitrust Litig.*, 985 F. Supp. 2d 1167 (N.D. Cal. 2013).

42. In settings where each observation in a sample represents a different unit of time—so called "time series" settings—there are statistical formulations of causality that rely on the ordering of events. *See* Pindyck & Rubinfeld, *supra* note 23, § 9.2. For a more general description of time series analysis, *see* James H. Stock & Mark W. Watson, Introduction to Econometrics, Chapter 15 (2019).

however, the information about the employer's use of the test scores could be translated into a second equation in which a new dependent variable, test score, is related to workers' wages and other variables. A test of the hypothesis that salary and race affect test scores would provide a suitable test of the absence of feedback. Finally, it might be determined that a variable that measures years of experience might be a suitable instrumental variable (i.e., it might be positively correlated with the test scores, yet directly unaffected by the dependent variable salary).[43]

When a suitable instrumental variable has been found, a more advanced form of multiple regression analysis known as instrumental variables regression will be appropriate.[44] Intuitively, with this form of regression analysis, the endogenous variable of interest is divided into two parts: (1) a component that is affected by the instrumental variable, but by assumption is not affected by feedback from the dependent variable (or by a spurious correlation); and (2) the remainder. An instrumental variable regression model only uses the part of the endogenous regressor that is affected by the instrumental variable. For more about the use of instrumental variables estimation, see the causal analysis discussion in the section titled "Causal Analysis and Research Designs," below.

Choosing Methods of Analysis Other Than the Basic Multiple Regression Method

There are many multivariate statistical techniques other than the basic multiple regression method that can be useful in legal proceedings. Some statistical methods are appropriate when nonlinearities are important,[45] while others are appropriate for models in which the dependent variable is discrete, rather than continuous.[46] Still others have been utilized predominantly to respond to methodological concerns arising in the context of discrimination litigation.[47]

43. Ideally, the instrumental variable should be uncorrelated with any sources of error that might diminish the measured effect of test scores on wages.

44. For examples of this practice in trial, see *United States v. Aetna, Inc.*, 240 F. Supp. 3d 1 (D.D.C. 2017) (antitrust action); *see also In re* Domestic Drywall Antitrust Litig., 322 F.R.D. 188 (E.D. Pa. 2017).

45. These techniques include, but are not limited to, piecewise linear regression, maximum likelihood estimation of models with nonlinear functional relationships, and autoregressive and moving-average time-series models.

46. For a general discussion of the probit model and its cousin the logit model, see Stock & Watson, *supra* note 42, at Chapter 11.

47. In the analysis of salary discrimination claims, some statisticians have suggested alternative approaches, including urn models (Bruce Levin & Herbert Robbins, *Urn Models for Regression Analysis, with Applications to Employment Discrimination Studies*, 46 Law & Contemp. Probs., 247 (1983)) and, as a means of correcting for measurement errors, reverse regression (Delores A.

It is essential that a valid statistical method be applied to assist with the analysis in each legal proceeding. Therefore, the expert should be prepared to explain why any chosen method, including multiple regression, was more suitable than the alternatives. The following discussion highlights several alternative methods that have proven useful when the dependent variable is discrete rather than continuous.

Suppose that a study has been offered into evidence that purports to evaluate whether there are racial disparities in the imposition of guilty verdicts by juries in criminal cases. One possible goal is to characterize the most important attributes of those cases in which juries find defendants guilty. One covariate might measure the severity of the crime and another might account for the race of the defendant. If the basic regression model is used, predictions generated by the regression model can be interpreted as a measure of the probability that a given defendant will be found to be guilty. Unfortunately, the basic regression model (the "linear probability model") leaves open the possibility that the predicted probabilities might lie below 0 or above 1—making the interpretation of the regression results nearly impossible.

A common solution to this problem is to utilize a probit or logit model in which the predicted values of the regression model are measures of a probability that lies within the 0 to 1 range. Probit[48] and logit[49] models are valuable tools in a wide range of empirical studies, but their results must be interpreted with care. For one thing, the effect of a change in the value of any of the covariates in the model on the probability of an outcome occurring (e.g., a high or a low probability of being found guilty) depends on the baseline probability in the absence of the change. If that probability is very high, a change in the covariate cannot raise the probability much further. Similarly, if the probability is very low, even a large change in the covariate may have a small effect. For example, in a probit or logit model for the probability a student is admitted to a selective college, a given

Conway & Harry V. Roberts, *Reverse Regression, Fairness, and Employment Discrimination*, 1 J. Bus. & Econ. Stat. 75 (1983), https://doi.org/10.2307/1391775). *But see* Arthur S. Goldberger, *Redirecting Reverse Regressions*, 2 J. Bus. & Econ. Stat. 114 (1984); Arlene S. Ash, *The Perverse Logic of Reverse Regression*, *in* Statistical Methods in Discrimination Litigation 85 (D.H. Kaye & Mikel Aickin eds., 1987).

48. Examples of probit analyses span several types of litigation. *See* Moussouris v. Microsoft Corp., 311 F. Supp. 3d 1223 (W.D. Wash. 2018) (sex discrimination); Chi. Teachers Union, Local 1 v. Bd. of Educ. of Chi., No. 12-C-10311, 2020 U.S. Dist. LEXIS 32351 (N.D. Ill. Feb. 25, 2020) (race discrimination); *In re* NCAA Athletic Grant-In-Aid Cap Antitrust Litig., No. 4:14-CV-02758, 2017 U.S. Dist. LEXIS 201104 (N.D. Cal. Dec. 6, 2017) (antitrust case); Georgia v. Ashcroft, 195 F. Supp. 2d 25 (D.D.C. 2002) (challenge to redistricting plans).

49. Logit models are likewise utilized in a variety of cases. *See* W.L. Gore & Assocs. v. C.R. Bard, Inc., No. 11-515-LPS-CJB, 2015 U.S. Dist. LEXIS 191654 (D. Del. Sept. 25, 2015) (patent infringement); Allegra v. Luxottica Retail N. Am., 341 F.R.D. 373 (E.D.N.Y. 2022) (false advertising class action); Laumann v. NHL, 117 F. Supp. 3d 299 (S.D.N.Y. 2015) (discussing "logit error" in a model in an antitrust violation case).

change in the student's SAT score will have little effect if the student already has an admission probability of 90%, or if the student has a probability of admission close to 0%. But it could have a large effect for students "on the bubble" with roughly a 50% chance of admission. This variation can be addressed by using the estimated model to calculate the change in the probability for each unit in the sample and then averaging the changes to get the average marginal effect of a change in the covariate.

For situations in which the dependent variable takes on only two values, logit and probit models are quite similar. They are also similar in cases where the dependent variable takes on ordered values (like 1, 2, 3, 4, 5)—for example, in situations where the dependent variable is a measure of agreement or sentiment recorded on a 5-point Likert scale (strongly agree, agree, neutral, disagree, strongly disagree).[50] But in cases where the dependent variable reflects three or more unordered values (e.g., an analysis of consumer demand for sedans, SUVs, and sports cars) a generalization of the logit model known as the multinomial logistic (MNL) model is particularly convenient. Furthermore, if a probabilistic interpretation is useful, the dependent variable can be specified as the logarithm of the odds that a particular choice will be made. In this special case, the coefficients of an estimated logit or MNL model will measure the logarithm of the odds of each of the various choices being made.

Interpretation of the results of probit and logit models can be difficult. To illustrate, assume that the logit model is being used to understand the factors that affect individuals who have recently moved into a city to buy (1) or not to buy (0) a house. Assume also that 60% of the individuals did indeed buy a house. In this case, it would be useful if the expert were asked to answer the following questions:

a. What are the measured effects of a change in each of the covariates of interest in the model on the probability that there will be a purchase (a 1)?
b. Are those calculated changes in probability evaluated for every individual in the sample and then averaged? If not, why not?
c. Do you have a measure of the "goodness of fit" of the logit model? To what extent does the logit model improve in terms of its predictive ability on a model that simply predicts which individuals will buy and which will rent at random (with a probability of purchase being 0.6)?

50. Such models—where the dependent variable involves three or more choices that have a natural ordering—are known as ordered probit and ordered logit regression models.

Model Selection

During testimony, an expert will often offer a regression model that is the result of an iterative model selection process. Depending on the issues in the case, that process might involve varying any or all of the following: (1) the choice of covariates in the model; (2) the form of the model (e.g., measuring the dependent variable as a logarithm or not, including interaction terms or not); (3) the selection of the sample to be used in estimating the model; and (4) the assumptions with respect to the underlying error structure of the model (e.g., accounting for a nonconstant error variance or the possibility that errors are correlated over time).[51]

A particular concern with model selection in legal settings is that the process of selection was driven in part by the desire of the expert to find a specification in which the coefficient of the variable of interest is either "statistically significant" or not. Such a process is commonly known as *p*-hacking because the statistical significance of an estimate is often expressed by its associated *p*-value.[52] Informally, this is the probability that the analyst would obtain the estimate in hand, even though the true underlying coefficient (i.e., the one that would be obtained using a very large sample) is zero. As discussed in the section titled "What Is the Appropriate Level of Statistical Significance?" below, it is conventional in scientific work to deem estimated coefficients with *p*-values of less than 5% as "statistically significant." Thus, an expert who is arguing that there is a "significant effect" of a variable of interest has an incentive to select a specification in which the *p*-value is 5% or smaller. An expert who is arguing the opposite—that the variable of interest has "no significant effect"—has an incentive to select a specification in which the *p*-value is above 5%. *P*-hacking is often suspected when the *p*-value of a selected model is just under or just over 5%, particularly when alternative specifications show much different levels of significance.

Finally, it is sometimes the case that the expert will state explicitly or rely implicitly on regression analyses and tests that were performed either by members of the expert's team or by other experts or teams. Reliance on the work of others should be reported as part of the expert's testimony. Otherwise, an expert could evade discovery obligations by having a consulting expert search for a model that the testifying expert then presents as if it were the expert's own.

The model searching exercise should be distinguished, at least conceptually, from the important process of testing a chosen regression model for robustness.

51. According to Peter Kennedy, "model specification should not blindly follow testing procedures . . . it needs to be a well-thought-out combination of theory and data, and . . . testing procedures used in such specification searches should be designed to minimize the costs of data mining." Peter E. Kennedy, *Sinning in the Basement: What Are the Rules? The Ten Commandments of Applied Econometrics*, 16 J. Econ. Surveys 569, 578 (2002), https://doi.org/10.1111/1467-6419.00179.

52. Megan L. Head et al., *The Extent and Consequences of* P-*Hacking in Science*, 13 PLoS Biology 1, 15 (2015), https://doi.org/10.1371/journal.pbio.1002106.

Robustness checks are applied once a model specification has been selected. The expert applies these checks to report the sensitivity of the reported results to alternative specifications.

The model searching process chosen by an expert may be ad hoc, that is, chosen to be appropriate for the specifics of the issue being studied. However, as computing power has increased, it has become more common for experts to use a variety of algorithmic approaches to estimate model parameters and make predictions. Algorithms can be especially useful when there are many covariates that are potential explanatory variables. Multiple regression is the most basic and most common algorithm, but there are a variety of other model-searching approaches that can benefit from the use of artificial intelligence (AI) methods. They include, but are not limited to, the following:

- Matching
- Least absolute shrinkage and selection operator (lasso) regression
- Random forest regression

Matching

Consider a case where the liability turns on the relationship between an employer's imposition of an anti-female policy (the variable of interest) and whether the individual is promoted or not (the dependent variable). Among all the available observations, some will be more informative than others about the nature of this relationship. To be specific, the most salient information might come by comparing (or matching) women to men of similarly observable characteristics (e.g., age, education, years of employment). There are a variety of algorithmic methods that are designed to find the best match from among all the possible potential matches in the sample of available data.

If one compares the promotion outcomes of women to those of men with similar ages, educational background, and years of experience, a regression model might not need to include those covariates. In its simplest form, the matching approach would measure the impact of the actions of the anti-female policy as the difference in the means of the promotion rate (the dependent variable) of those adversely affected and those not for the reduced sample of matched observations. In most cases there will not be exact matches—hence the need for an algorithmic approach that chooses a subset of the observations likely to be most informative, or more generally weighting the observations from the comparison group in some manner.[53]

53. Weighting methods are discussed in detail in Nicole Fortin, Thomas Lemieux & Sergio Firpo, *Decomposition Methods in Economics*, *in* 4 Handbook of Labor Economics, Part A 1–102 (Orley Ashenfelter & David Card eds., 2011).

The matching approach has the advantage that it does not require the expert to specify the form of the underlying regression model—it could be linear or log-linear, for example, or it could include interaction variables. However, matching procedures may end up using a small fraction of the available data from the comparison group, particularly if the group of interest (e.g., female employees) constitutes a small share of the overall sample. In some cases, this loss of sample size may lead to imprecision in the comparison of interest. In addition, matching procedures will generate results of which the significance is not easily tested statistically. Finally, different matching algorithms have different options (sometimes called tuning parameters) that must be set, specifying the tradeoffs to be made in evaluating potential matches and in deciding whether to retain observations when no close match can be found. Since the expert has discretion to choose among these options, it is important for the expert to explain how the choices were made.

Least absolute shrinkage and selection operator (lasso) regression

Lasso can be a useful methodology when there are many covariates that are highly correlated with each other—a situation referred to as multicollinearity. Lasso reduces the impact of multicollinearity by selecting a smaller set of covariates for predictive use in a regression context. If effective, lasso regression can reduce the possible instability of regression models, and it has the potential to improve the prediction process.[54]

Random forest regression

The random forest regression approach is a decision-tree modeling approach that divides the data into two or more parts, searches among possible regression specifications for each part, and then merges the predictions for each part to (ideally) generate a more accurate and stable prediction that would arise if a more basic multiple regression methodology were utilized.[55] Relying on decision trees can be beneficial when a study involves a large dataset and when one believes that there is a nonlinear and/or highly interactive relationship between the covariates and the dependent variable. The random forest approach is less likely to be of value when the data are

54. Unlike ordinary least squares (OLS), which minimizes the sum of squared residuals, the Lasso estimator minimizes the sum of squares of the values of the coefficients in a model in which all of the variables are specified so that the coefficients are comparable. *See* Stock & Watson, *supra* note 42, § 14.4. Similarly, a related technique, ridge regression, reduces the number of covariates by utilizing a penalty that increases with the sum of the squared coefficients. *See id.* § 14.3.

55. Of note, the bootstrapping random forest method algorithm creates randomly drawn decision trees from the data, and then averages the results. For a description of the statistical foundation of random forest regression, see Leo Breiman, *Random Forests*, 45 Mach. Learning 5–32 (2001).

in the form of a time series. Time series modeling and analysis often require that one utilize specialized methods to create and then analyze a stationary data series.

Interpreting Multiple Regression Results

Multiple regression results can be interpreted in purely statistical terms, through significance tests, or they can be interpreted in a more practical, nonstatistical manner. Although an evaluation of the practical significance of regression results is almost always relevant in the courtroom, tests of statistical significance are appropriate only in particular circumstances.

What Is the Practical, as Opposed to the Statistical, Significance of the Regression Results?

Practical significance means that the magnitude of the effect being studied is not de minimis (i.e., it is sufficiently important for the court to be concerned). For example, if the average weekly wage rate is $200, a wage differential between men and women of $2 is likely to be deemed practically insignificant because the differential represents only 1% of the average weekly wage.[56] That same difference could be statistically significant, however, if a sufficiently large sample of men and women were studied.[57] The reason is that statistical significance is determined, in part, by the number of observations in the dataset.

Often, results that are practically significant are also statistically significant.[58] However, it is possible with a large dataset to find statistically significant

56. There is no specific percentage threshold above which a result is practically significant. Practical significance must be evaluated in the context of a particular legal issue. *See also* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "*p*-values, Significance Levels, and Hypothesis Tests" section, in this manual.

57. Practical significance also can apply to the overall credibility of the regression results. Thus in *McCleskey v. Kemp*, 481 U.S. 279 (1987), coefficients on race variables were statistically significant, but the Court declined to find them legally or constitutionally significant.

58. In *Melani v. Board of Higher Education*, 561 F. Supp. 769, 774 (S.D.N.Y. 1983), a Title VII suit was brought against the City University of New York (CUNY) for allegedly discriminating against female instructional staff in the payment of salaries. One approach of the plaintiff's expert was to use multiple regression analysis. The coefficient on the variable that reflected the sex of the employee was approximately $1,800 when all years of data were included. Practically (in terms of average wages at the time) and statistically (in terms of a 5% significance test), this result was significant. Thus, the court stated, "Plaintiffs have produced statistically *significant* evidence that women hired as CUNY instructional staff since 1972 received *substantially* lower salaries than similarly

coefficients that are practically insignificant. On the other hand, it is also possible to obtain results that are practically significant but fail to achieve statistical significance (especially when the sample size is small). Suppose for example that an expert undertakes a damages study in a patent-infringement case and predicts but-for sales—what sales would have been had the infringement not occurred—using data that predate the period of alleged infringement. If data limitations are such that only three or four years of pre-infringement sales are known, the difference between but-for sales and actual sales during the period of alleged infringement could be practically significant but statistically insignificant. Alternatively, with only three or four data points, the expert will be unable to detect an effect, even if one exists.

When Should Statistical Tests Be Used?

A test of a specific contention, a hypothesis test, often assists the court in determining whether a violation of the law has occurred in areas in which direct evidence is inaccessible or inconclusive. For example, an expert might use hypothesis tests in race or sex discrimination cases to determine the presence of a discriminatory effect.

Statistical evidence alone can never prove with absolute certainty the worth of any substantive theory. However, by providing evidence contrary to the view that a particular form of discrimination has not occurred, for example, the

qualified men." *Id*. at 781 (emphasis added). For a related analysis involving multiple comparison, see *Csicseri v. Bowsher*, 862 F. Supp. 547, 572 (D.D.C. 1994) (noting that plaintiff's expert found "statistically significant instances of discrimination" in 2 of 37 statistical comparisons, but suggesting that "2 of 37 amounts to roughly 5% and is hardly indicative of a pattern of discrimination"), *aff'd*, 67 F.3d 972 (D.C. Cir. 1995). *See also* Apsley v. Boeing Co., 722 F. Supp. 2d 1218 (D. Kan. 2010) (writing that "[s]tatistical significance does not tell us whether the disparity we are observing is meaningful in a practical sense nor what may have caused the disparity" and finding that plaintiffs did not establish a prima facie case that if "forty-eight more people over the age of 40 would have been hired, Plaintiffs' hiring statistics would not have been statistically significant"). Practical significance continues to be important in other contexts. It often arises in disparate impact cases in the form of a disputed but widely used "four-fifths rule," an EEOC rule of thumb that if the hire or promotion rate of a minority group is four-fifths that of the comparison group (often white employees), the difference is practically significant. *See, e.g.*, Hispanic Nat'l L. Enf't Ass'n NCR v. Prince George's Cnty., 535 F. Supp. 3d 393 (D. Md. 2021) (writing that "pairing [of] a statistical significance test with a practical significance measure is the contemporary standard for demonstrating disparate impact") (citation omitted); *cf.* Jones v. City of Boston, 752 F.3d 38, 53 (1st Cir. 2014) (concluding that "a plaintiff's failure to demonstrate practical significance cannot preclude that plaintiff from relying on competent evidence of statistical significance to establish a prima facie case of disparate impact").

multiple regression approach can aid the trier of fact in assessing the likelihood that discrimination has occurred.[59]

Tests of hypotheses are appropriate in a cross-sectional analysis, in which the data underlying the regression study have been chosen as a sample of a population at a particular point in time, and in a time-series analysis, in which the data being evaluated cover several time periods. In either analysis, the expert may want to evaluate a specific hypothesis, usually relating to a question of liability or to the determination of whether there is measurable impact of an alleged violation. Thus, in a sex discrimination case, an expert may want to evaluate a null hypothesis of no discrimination against the alternative hypothesis that discrimination takes a particular form.[60] In this case, it is important to realize that rejection of the null hypothesis does not in itself prove legal liability. It is possible to reject the null hypothesis and believe that an alternative explanation other than one involving legal liability accounts for the results.[61]

What Is the Appropriate Level of Statistical Significance?

In most scientific work, the level of statistical significance required to reject the null hypothesis (i.e., to obtain a statistically significant result) is set conventionally at 0.05, or 5%.[62] The significance level measures the probability that the null hypothesis will be rejected incorrectly. In general, the lower the percentage required for statistical significance, the more difficult it is to reject the null hypothesis, and therefore it is less likely that one will err in doing so. Although the 5%

59. *See* Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324 (1977) (the Court inferred discrimination from overwhelming statistical evidence by a preponderance of the evidence); Ryther v. KARE 11, 108 F.3d 832, 844 (8th Cir. 1997) ("The plaintiff produced overwhelming evidence as to the elements of a prima facie case, and strong evidence of pretext, which, when considered with indications of age-based animus in [plaintiff's] work environment, clearly provide sufficient evidence as a matter of law to allow the trier of fact to find intentional discrimination."); Paige v. California, 291 F.3d 1141 (9th Cir. 2002) (allowing plaintiffs to rely on aggregated data to show employment discrimination).

60. Tests are also appropriate when comparing the outcomes of a set of employer decisions with those that would have been obtained had the employer made a different choice from among the available options.

61. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Evaluating Hypothesis Tests" in this manual, and see the section titled "Difference-in-Differences Design" below.

62. *See, e.g.*, Palmer v. Shultz, 815 F.2d 84, 92 (D.C. Cir. 1987) ("'the .05 level of significance . . . [is] certainly sufficient to support an inference of discrimination'" (quoting Segar v. Smith, 738 F.2d 1249, 1283 (D.C. Cir. 1984), *cert. denied*, 471 U.S. 1115 (1985))); United States v. Delaware, Civil Action No. 01-020-KAJ, 2004 U.S. Dist. LEXIS 4560 (D. Del. Mar. 22, 2004) (stating that .05 is the normal standard chosen).

criterion is typical, reporting of more stringent 1% significance tests or less stringent 10% tests can also provide useful information.

In conducting a statistical test, it is useful to compute an observed significance level, or *p*-value. The *p*-value associated with the null hypothesis that a regression coefficient is 0 is the probability that a coefficient of this magnitude or larger could have occurred by chance if the null hypothesis were true. If the *p*-value were less than or equal to 5%, the expert would reject the null hypothesis in favor of the alternative hypothesis, whereas if the *p*-value were greater than 5%, the expert would fail to reject the null hypothesis. The use of 1%, 5%, and sometimes 10% levels for determining statistical significance remains a subject of debate. One might argue, for example, that when there is a relatively specific alternative to the null hypothesis, a somewhat lower level of confidence might be appropriate. Otherwise, when the alternative to the null hypothesis involves a vague alternative of "effect," a high level of confidence (associated with a low significance level, such as 1%) may be appropriate.[63]

Should Statistical Tests be One-Tailed or Two-Tailed?

When the expert evaluates the null hypothesis that a variable of interest has no linear association with a dependent variable against the alternative hypothesis that there is an association, a two-tailed test, which allows for the effect to be either positive or negative, is usually appropriate. A one-tailed test would usually be applied when the expert believes, perhaps based on other direct evidence presented at trial, that the alternative hypothesis is either positive or negative, but not both. For example, an expert might use a one-tailed test in a patent infringement case if they strongly believe that the effect of the alleged infringement on the price of the infringed product was either zero or negative. (The sales of the infringing product competed with the sales of the infringed product, thereby

63. *See, e.g.*, Vuyanich v. Republic Nat'l Bank, 505 F. Supp. 224, 272 (N.D. Tex. 1980) (noting the "arbitrary nature of the adoption of the 5% level of [statistical] significance" to be required in a legal context); Cook v. Rockwell Int'l Corp., 580 F. Supp. 2d 1071 (D. Colo. 2006). Indeed, "courts generally have found that challenges to statistical significance go to the weight, but not the admissibility, of a regression model." *In re* EpiPen (Epinephrine Injection, USP) Mktg., Sales Pracs. and Antitrust Litig., No. 17-MD-2785-DDC-TJJ, 2020 U.S. Dist. LEXIS 40788, at *131 (D. Kan. Feb. 27, 2020). *See* Kurtz v. Kimberly-Clark Corp., 414 F. Supp. 3d 317, 331 (E.D.N.Y. 2019) ("Regressions should not be excluded on the ground that they fail to meet arbitrary thresholds of statistical significance."); *In re* High-Tech Emp. Antitrust Litig., No. 11-CV-02509-LHK, 2014 WL 1351040, at *15 (N.D. Cal. Apr. 4, 2014) ("The Court finds that the fact that these two variables are not statistically significant at the 1%, 5%, and 10% levels goes to the weight, not the admissibility of [the] model."); EEOC v. Mavis Discount Tire, Inc., 129 F. Supp. 3d 90, 110 (S.D.N.Y. 2015) (writing that while courts have rejected a "formal" litmus test for Title VII claims, two or three standard deviations "can be highly probative") (citation omitted).

lowering the price.) By using a one-tailed test, the expert is in effect stating that without analyzing the data, it would be very surprising if the data pointed in the direct opposite direction to the one posited by the expert.

Because using a one-tailed test produces *p*-values that are one-half the size of *p*-values using a two-tailed test, the choice of a one-tailed test makes it easier for the expert to reject a null hypothesis. Correspondingly, the choice of a two-tailed test makes null hypothesis rejection less likely. Because there is some arbitrariness involved in the choice of an alternative hypothesis, courts should avoid relying solely on sharply defined statistical tests.[64] However, reporting the *p*-value or a confidence interval should be encouraged because it conveys useful information to the court, whether a null hypothesis is rejected or not.

Are the Regression Results Robust?

The issue of robustness—whether regression results are sensitive to slight modifications in assumptions (e.g., that the data are measured accurately)—is of vital importance. If the assumptions of the regression model are valid, standard statistical tests can be applied. When the assumptions of the model are violated, standard tests can overstate or understate the significance of the results.

The violation of an assumption does not necessarily invalidate a regression analysis, however. The following questions highlight some of the more important assumptions of regression analysis.

64. Courts have shown a preference for two-tailed tests. *See, e.g.*, Palmer v. Shultz, 815 F.2d 84, 95–96 (D.C. Cir. 1987) (rejecting the use of one-tailed tests, the court found that because some appellants were claiming over-selection for certain jobs, a two-tailed test was more appropriate in Title VII cases); Smith v. City of Boston, 144 F. Supp. 3d 177, 196 (D. Mass. 2015) (observing in a Title VII case, "[t]he weight of the case law appears to favor two-tailed tests"); Moore v. Summers, 113 F. Supp. 2d 5, 20 (D.D.C. 2000) (reiterating the preference for a two-tailed test). *See also* Csicseri v. Bowsher, 862 F. Supp. 547, 565 (D.D.C. 1994) (finding that although a one-tailed test is "not without merit," a two-tailed test is preferable); *but see In re* EpiPen (Epinephrine Injection, USP) Mktg., Sales Pracs. and Antitrust Litig., No. 17-MD-2785-DDC-TJJ, 2020 U.S. Dist. LEXIS 40788, at *131 (D. Kan. Feb. 27, 2020) (allowing a model with a one-tailed test to move to the factfinder, writing that the expert's reasoning was "at worst, plausible"). *See also* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Evaluating Hypothesis Tests" in this manual, and see also the section titled "Difference-in-Differences Design" below.

What Evidence Exists That the Explanatory Variable Exerts a Causal Effect on the Dependent Variable and Is Not Affected by Endogeneity?

In a multiple regression framework, the expert often assumes that changes in one or more of the explanatory variables affect the dependent variable, but changes in the dependent variable do not affect the explanatory variables (i.e., no feedback, as described previously), nor is there any spurious correlation arising from unmeasured factors that affect both variables. If the causality were reversed, or if there were unmeasured factors leading to spurious correlation, the expert and the trier of fact would likely reach the wrong conclusion from the reported results.

As noted in the section titled "Choosing the Additional Explanatory Variables" above, there are no basic, direct statistical tests for determining the direction of causality. Rather, it may be appropriate to ask (1) whether there is any concern over feedback (or reverse causality) from the outcome variable to the explanatory variable; and (2) whether there is any concern over possible unmeasured factors that affect both variables.

If there is a possibility of feedback from the outcome variable to one or more of the explanatory variables in the regression model, it may be useful to ask the following questions:

a. How do the results change if the questionable explanatory variable(s) are dropped from the model?
b. Has the expert developed a model relating the questionable variable to the outcome variable and the other covariates? If so, is there evidence of feedback from the outcome variable to the questionable variable?

To What Extent Are the Explanatory Variables Correlated with Each Other?

It is essential in multiple regression analysis that the explanatory variable of interest not be correlated perfectly with one or more of the other explanatory variables. In essence, with perfect correlation there are two explanations for the same pattern in the data. Suppose, for example, that in a gender discrimination suit, a particular form of job experience is determined to be a valid source of high wages. If all men had the requisite job experience and all women did not, it would be impossible to tell whether wage differentials between men and women resulted from gender discrimination or simply differences in years of experience.

When two or more explanatory variables are correlated perfectly—that is, when there is perfect collinearity—one cannot estimate the regression parameters. The existing dataset does not allow one to distinguish between

alternative competing explanations of the movement in the dependent variable. However, when two or more variables are highly, but not perfectly, correlated—that is, when there is multicollinearity—the regression can be estimated, but some concerns remain. As discussed above, the greater the multicollinearity between two variables, the less precise are the estimates of individual regression parameters, and the less an expert is able to distinguish among competing explanations for the movement in the outcome variable (even though there is no problem in estimating the joint influence of the two variables and all other regression parameters).[65]

Fortunately, the reported regression statistics take into account any multicollinearity that might be present.[66] However, it is important to note as a corollary that a failure to find a strong relationship between a variable of interest and a dependent variable need not imply that there is no relationship.[67] A relatively small sample, or even a large sample with substantial multicollinearity, may not provide sufficient information for the expert to determine whether there is a relationship.

65. *See* Griggs v. Duke Power Co., 401 U.S. 424 (1971) (The court argued that an education requirement was one rationalization of the data, but racial discrimination was another. Putting both race and education in the regression, it would have been asking too much of the data to tell which variable was doing the real work, because education and race were so highly correlated in the market at that time.).

66. *See* Denny v. Westfield State College, 669 F. Supp. 1146, 1149 (D. Mass. 1987) (The court accepted the testimony of one expert that "the presence of multicollinearity would merely tend to *overestimate* the amount of error associated with the estimate. In other words, *p*-values will be artificially higher than they would be if there were no multicollinearity present.") (emphasis added); *In re* High Fructose Corn Syrup Antitrust Litig., 295 F.3d 651, 659 (7th Cir. Ill. 2002) (refusing to second-guess district court's admission of regression analyses that addressed multicollinearity in different ways). Multicollinearity has not been a reason to preclude a model from the factfinder. *See In re* Air Cargo Shipping Servs. Antitrust Litig., No. 06-MD-1175, 2014 WL 7882100, at *19 (E.D.N.Y. Oct. 15, 2014), *report and recommendation adopted*, 2015 WL 5093503 (E.D.N.Y. July 10, 2015) ("The fact that Kaplan's model may be tainted by multicollinearity goes purely to its weight, and is not a reason to strike it."); *In re* High-Tech Employee Antitrust Litig., No. 11-CV-02509, 2014 WL 1351040, at *21 (N.D. Cal. Apr. 4, 2014) ("Other courts have admitted regressions even in the face of expert disagreement regarding whether collinearity posed a problem. . . . This is not surprising given that the concept of collinearity is not a methodology, but a common phenomenon that results when using the methodology of regression analysis."); *In re* Korean Ramen Antitrust Litig., No. 13-CV-04115-WHO, 2017 U.S. Dist. LEXIS 7756 (N.D. Cal. Jan. 19, 2017) (finding a permissible level of multicollinearity and dispute over variable inclusion did not make model unreliable). *But see* Reed Const. Data Inc. v. McGraw-Hill Cos., Inc., 49 F. Supp. 3d 385, 405 (S.D.N.Y. 2014), *aff'd*, 638 F. App'x 43 (2d Cir. 2016) (excluding expert's method, in part, because of severe multicollinearity).

67. If an explanatory variable of concern and another explanatory variable are highly correlated, dropping the second variable from the regression can be instructive. If the coefficient on the explanatory variable of concern becomes significant, a relationship between the dependent variable and the explanatory variable of concern is suggested.

How Is the Sample Used in the Regression Model Defined?

One potential source of bias in regression modeling arises when the sample used to estimate the model represents only a subsample of the underlying population, and inclusion in the sample is based on a criterion that may be related to the outcome or to the explanatory variables in the proposed regression model. For example, consider an analysis of possible racial disparities in the imposition of guilty verdicts by juries in criminal cases. Such an analysis can only be conducted for cases that go to trial. But if attorneys for nonwhite defendants are aware of potential jury bias, they may recommend a plea bargain unless they believe the case is highly unlikely to yield a guilty verdict despite the race of the defendant. In that situation, the set of cases with nonwhite defendants that appear at trial will tend to be highly selective, and less likely to have a guilty verdict than other cases, leading to a sample selection bias in the estimated effect of defendant race on the probability of a guilty verdict.[68] Sample selection bias can also arise even when an initial sample is randomly selected, if inclusion in the final sample used in the model estimation depends on the presence of certain information (for example, the availability of complete data on key variables).

The expert should be prepared to explain the derivation of the sample used in estimation. In situations where the estimation sample differs from the underlying population and is not based on a random sample selection rule, the expert should be prepared to defend the sample selection criteria and explain why these criteria do not lead to sample selection bias.

Are There Problems with Statistical Inference Owing to Nonindependent Errors?

Multiple regression analysis yields a set of estimated coefficients that measure the effects of each of the explanatory variables on the outcome variable, holding constant the other explanatory variables. Standard regression modeling software also reports a standard error for each coefficient that can be used to form a *t*-statistic or a confidence interval.[69] To this point we have emphasized problems such as endogeneity that can arise in interpreting these estimated coefficients from a regression model. But a second set of problems can arise in estimating the standard

68. Sample selection bias is also an issue for comparisons of the outcomes of police searches of motorists, since only motorists who are stopped in the first place are searched. *See, e.g.*, John Knowles, Nicola Persico & Petra Todd, *Racial Bias in Motor Vehicle Searches: Theory and Evidence*, 109 J. Pol. Econ. 203 (2001).

69. For a discussion of confidence intervals, see Stock & Watson, *supra* note 42, at Chapter 3. Loosely speaking, a confidence interval represents an interval of values in which the true value of a regression coefficient falls within some pre-specified probability (where the true value is the estimate that would be obtained from the same model with a very large sample).

errors, even when the coefficient estimates are unbiased. An important assumption underlying the procedure used to estimate these standard errors is that the error terms in the regression model are independent of each other and independent of the explanatory variables.[70] Independence of the errors is a much stronger assumption than is needed to ensure unbiasedness of the coefficient estimates.

The assumption of independence may be inappropriate in several circumstances. If this strong assumption does not hold and the estimates of the standard errors are themselves biased, the analyst might draw inappropriate conclusions about the confidence that should be attributed to the size of the regression coefficients. In some instances, failure of the assumption makes multiple regression analysis an unsuitable statistical technique; in other instances, modifications or adjustments within the regression framework can be made to accommodate the failure.

The independence assumption may fail, for example, in a study of individual behavior over time, in which an unusually high error value in one period is likely to lead to an unusually high value in the next period. For example, if an economic forecast model underpredicted this year's gross domestic product, the model is likely to underpredict next year's as well; the factor that caused the prediction error (e.g., an incorrect assumption about Federal Reserve policy) is likely to be a continuing source of error in the future.[71]

Alternatively, the assumption of independence may fail in a study of a group of firms at a particular point in time in which the error terms for large firms are systematically larger in absolute value than the error terms for small firms.[72] For example, an analysis of the profitability of firms will include in the error term all the unaccounted-for factors that lead to higher or lower profit. For a large national retail firm with many stores, these omitted factors can lead to errors of plus or minus several million dollars in magnitude. For a small local retailer with one store, however, the range of omitted factors is smaller, and the errors may only be plus or minus a few thousand dollars.

A third possibility is that the dependent variable varies at the individual level, but the explanatory variable of interest varies only at the level of a group. For example, an expert might be viewing the price of a product in an antitrust case as a function of a variable or variables that depend on the marketing channel through which the product is sold (e.g., wholesale or retail). In this case, errors

70. When two variables are independent there is no information in one variable about any feature of the other variable. Independence implies that the variables are uncorrelated, which only requires that there is no information in one variable about the mean value of the other.

71. In this case, the errors in the regression model are said to be serially correlated.

72. This general class of problems, in which the variability in the error term is related to the covariates, is known as conditional heteroscedasticity. This problem also arises when the dependent variable is a binary variable (i.e., with a value of 0 or 1), because the range that the error term can take is limited.

within each of the marketing groups are not likely to be independent. Failure to account for this could cause the expert to overstate the statistical significance of the regression parameters.

In some instances, there are statistical tests that are appropriate for evaluating the assumption that the error terms are independent of each other and of the covariates in the model.[73] If the assumption has failed, there are more advanced versions of the procedures to estimate standard errors that can be used to correct for correlation in the errors over time, or for differences in the dispersion in the error component that depend on the covariates, or for dependence between observations from related subgroups of the dataset.[74] In some cases it is also possible to estimate an alternative version of the regression model that takes into account the dependence of the error terms for different observations.[75]

To What Extent Are the Regression Results Sensitive to Individual Data Points?

Evaluating the robustness of multiple regression results is a complex endeavor. Consequently, there is no agreed upon set of tests for robustness that analysts should apply. In general, it is important to explore the reasons for unusual data points. If the source is an error in recording data, the appropriate corrections can be made. If all the unusual data points have certain characteristics in common (e.g., they all are associated with a supervisor who consistently gives high ratings in an equal pay case), the regression model should be modified appropriately.

73. In a time-series analysis, the correlation of error values over time, the serial correlation, can be tested (in most instances) using several tests, including the Durbin-Watson test. The possibility that some error terms are consistently high in magnitude and others are systematically low—a form of heteroscedasticity—can also be tested in several ways. *See, e.g.*, Pindyck & Rubinfeld, *supra* note 23, at 146–59.

74. When serial correlation and/or heteroscedasticity are present, the standard errors associated with the estimated coefficients must be modified. For a discussion of the use of such "robust" standard errors, see Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, Chapter 8 (4th ed. 2009). For a discussion of the treatment of standard errors when the analysis involves panel data (cross-sections and time series), see Stock & Watson, *supra* note 42, at Chapter 10.

75. When serial correlation is present, several closely related statistical methods are appropriate, including generalized differencing (a type of generalized least squares) and maximum likelihood estimation. When heteroscedasticity is the problem, weighted least squares and maximum likelihood estimation are appropriate. *See* Stock & Watson, *supra* note 42, at Chapter 16. All these techniques are readily available in a variety of statistical computer packages. They also allow one to perform the appropriate statistical tests of the significance of the regression coefficients.

One might think that the sensitivity of regression results can be readily evaluated through an analysis of the regression residuals.[76] Unfortunately, this is not *always* the case. Given that the basic regression model penalizes estimated prediction errors by the square of the error, an outlier, a highly unusual data point that is more than some appropriate distance from a regression line (even if measured incorrectly), may affect the regression line very substantially, with the result being a relatively low regression residual.

To test to see if this is a problem, one useful diagnostic technique is to determine to what extent the estimated parameter changes as each data point in the regression analysis is dropped from the sample. An influential data point—a point that causes the estimated parameter to change substantially—should be studied further to determine whether mistakes were made in the use of the data or whether important explanatory variables were omitted.[77]

One final cautionary note: What is or is not an outlier is subjective. This leaves the possibility that an expert might point to an outlier or outliers in support of a particular point of view.

To What Extent Are the Data Subject to Measurement Error?

In multiple regression analysis it is assumed that variables are measured accurately.[78] If there are measurement errors in the dependent variable, estimates of regression parameters will be less precise, but they will not necessarily be biased.[79] However, if one or more independent variables are measured with error, the corresponding parameter estimates are likely to be biased, typically toward zero (and other coefficient estimates are likely to be biased as well). As the measurement error increases, the estimated parameter associated with the noisily measured

76. A regression residual is the difference between the actual value of the dependent variable and the value that is predicted by the estimated regression model.

77. A more complete and formal treatment of the robustness issue appears in David A. Belsley et al., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* 229–44 (1980). For a useful discussion of the detection of outliers and the evaluation of influential data points, see R.D. Cook & S. Weisberg, *Residuals and Influence in Regression* (Monographs on Stat. and Applied Probability No. 18, 1982). For a broad discussion of robust regression methods, see Peer J. Rousseeuw & Annick M. Leroy, *Robust Regression and Outlier Detection* (2003). *See generally* Peter J. Huber & Elvezio M. Rochetti, Robust Statistics (2d ed. 2009).

78. Inaccuracy can occur not only in the precision with which a particular variable is measured, but also in the precision with which the variable to be measured corresponds to the appropriate theoretical construct specified by the regression model.

79. An exception to this rule arises when the dependent variable is dichotomous (or takes on a discrete set of values). *See* Jerry Hausman, *Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left*, 15 J. Economic Perspectives 57, 67 (2001), https://doi.org/10.1257/jep.15.4.57.

variable will tend toward 0, that is, eventually there will be no relationship with the dependent variable.

It is important for any source of measurement error to be carefully evaluated. In some circumstances, little can be done to correct the measurement-error problem, and the regression results must be interpreted in that light. In other circumstances, however, the expert can correct measurement error by finding a new, more reliable data source. Finally, alternative estimation techniques (using related variables that are measured without error) can be applied to remedy the measurement-error problem in some situations.[80]

Causal Analysis and Research Designs

Multiple regression is a powerful tool for measuring the association between a specific variable or covariate of interest, such as employee gender, and an outcome, such as salary, while holding constant the effects of other observed variables. In policy analysis and in litigation there are many situations in which an expert will put forward a regression model and argue that the estimated regression coefficient associated with the variable of interest can be interpreted as an estimate of the causal effect of that variable, that is, the average change in the outcome that would occur if one were able to change the value of the variable of interest without changing *anything else*. This interpretation requires that all the possible factors that affect the outcome are either (1) included as covariates in the model, or (2) known to be uncorrelated with the variable of interest, so their omission from the model does not lead to bias in the estimated coefficient of the variable of interest.

Causal research designs are techniques to isolate the causal effect of a variable of interest in situations where some determinants of the outcome are omitted, and likely to be correlated with the variable of interest (a problem of omitted variables); or where the variable of interest is itself partly determined by the same unobserved factors that also affect the outcome (a problem of endogeneity).[81] In either case the essential idea is to isolate some part of the variation in the variable of interest that is arguably uncorrelated with the unobserved factors, and to focus on measuring the effect of that more limited variation on the outcome variable.

An illustrative example will make this framework clear. One researcher considers the question of whether an increase in the share of retail outlets owned by

80. *See, e.g.*, Pindyck & Rubinfeld, *supra* note 23, at 178–98 (discussion of instrumental variables estimation).

81. These two cases both give rise to a potential correlation between the variable of interest and the residual term in a basic regression model.

vertically integrated petroleum companies leads to higher gasoline prices.[82] Such an analysis could be presented as part of an argument that a proposed sale of gas stations owned by an independent retailer to a national company should be blocked, based on what had happened in earlier cases. Unfortunately, a simple regression model relating average gasoline prices (the outcome variable) to the share of vertically integrated sellers (a variable of interest) will not necessarily yield causally interpretable estimates, since the fraction of gas stations operated by vertically integrated sellers, as opposed to independent "non-branded" sellers, may be correlated with unobserved factors not taken into consideration by the statistical model. The author uses a difference-in-differences (DD) design that uses multiple regression to measure the effect of a change in the presence of non-branded sellers in specific market areas on the change in average gasoline prices in those areas, relative to price changes in other areas where there has been no change in the presence of non-branded sellers. She argues that price trends in the other areas are a good benchmark for how prices in the areas affected by the change would have otherwise evolved. In that case, deviations from this benchmark (which is what are measured in the difference-in-differences design) represent the causal effect of the change.

As another example, authors of another study ask how pretrial detention affects the case outcomes and future employment prospects of defendants.[83] As with the previous example, it is an open question as to whether the observed relationship between pretrial detention (the variable of interest) and defendant outcomes has a causal interpretation: defendants who receive more lenient bail terms are presumably different from those who do not. To derive a causal estimate, the authors use an instrumental variables design, in which the share of *other* defendants released prior to their trial by the bail judge is used as an instrumental variable for whether the defendant is detained pretrial or not. Since bail judges are usually randomly assigned, the authors argue that average judgments in other cases reflect judge-specific leniency, rather than features of the defendant or case. As a result, their design isolates differences in pretrial release rates, attributable to the luck of the draw in facing a harsh or lenient bail judge, that can be used to measure causal effects on subsequent case outcomes and employment rates.

The subsections that follow describe a range of methodological approaches that can and have been utilized to make causal inferences.

82. Justine S. Hastings, *Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California*, 94 Am. Econ. Rev. 317 (2004), https://doi.org/10.1257/000282804322970823.

83. Will Dobbie, Jacob Goldin, & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 Am. Econ. Rev. 201 (2018), https://doi.org/10.1257/aer.20161503.

Randomized Controlled Trials

The benchmark for causal designs is a simple randomized controlled trial (RCT, also known as a randomized experiment), as is widely used to evaluate new drugs and is sometimes used to analyze novel social programs such as an unemployment benefit for newly released prisoners.[84] In the simplest RCT, the analyst randomly divides potential subjects into two groups: the treatment group who receive the intervention (for example, the new drug or the unemployment benefits) and the control group who do not.[85] Because assignment to the two groups is random, the outcomes of the control group provide a valid basis for inferring what the outcomes for the treatment group would have been but for the effects of the treatment, that is, a valid counterfactual. As a result, any systematic differences between the treatment and control groups can be attributed to the intervention, without concern for the presence of omitted or unmeasured variables. An RCT solves the omitted variables problem by assigning the variable of interest to different subjects in a random way, thereby allowing one to assume that all factors that could influence the outcome are equally likely (or "balanced") in the treatment and control groups, even if these factors cannot be observed.[86]

A widely used conceptual framework for understanding the precise meaning of causality and the power of an RCT to measure the causal effect of an

84. The 1976 Transitional Aid Research Project was a randomized controlled trial in which some newly released prisoners received time-limited weekly benefits while they were out of work. Peter H. Rossi, Richard A. Berk & Kenneth J. Lenihan, Money, Work, and Crime: Experimental Evidence (1980). It followed an earlier randomized trial of a similar benefit program for newly released prisoners, known as the Baltimore Living Insurance for Ex-Prisoners project. Charles D. Mallar & Craig V. D. Thornton, *Transitional Aid for Released Prisoners: Evidence from the Life Experiment*, 13 J. Hum. Rsch. 208 (1978). Other applications include a 1990s U.S. Census experiment to find whether asking about Social Security numbers would decrease response rates—it did, finding a "3.4% decline in self-response rates attributable to the question." New York v. United States DOC, 351 F. Supp. 3d 502, 526 (S.D.N.Y. 2019). RCTs have been crucial for FDA approval when marketing certain products, including e-cigarettes. Wages & White Lion Invs., L.L.C. v. FDA, 41 F.4th 427 (5th Cir. 2022) (upholding FDA decision as neither arbitrary nor capricious, as plaintiffs did not bring sufficient safety evidence in the form of an RCT or longitudinal study).

85. In some cases there will be more than two treatment groups. For example, in the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984), police officers responding to misdemeanor domestic assaults were randomly assigned to one of three protocols: arrest the suspected offender; offer advice and mediation to the victim and offender; or send the suspect away for eight hours. Strictly speaking this experiment did not have a control group, but any one of the treatment groups can be compared against the other two. Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 Am. Socio. Rev. 261 (1984), https://doi.org/10.2307/2095575.

86. One federal court assessed RCT as the strongest type of evidence: "Scientific evidence is assessed on a scale of Level I to Level V, with Level I granted the most scientific weight and Level V the least. A randomized controlled trial is an example of Level I evidence, expert opinion is Level V." McClellan v. I-Flow Corp., 710 F. Supp. 2d 1092, 1107 n.10 (D. Or. 2010) (citation omitted).

intervention was proposed by Donald Rubin in 1974.[87] In this framework, we hypothesize that each subject has two potential outcomes: one if they were to receive the intervention, another if they did not. For example, patients interested in receiving blood pressure medication over the next year would have one blood pressure reading at the end of the year if they were to take the drug, but a different reading if they did not take the drug. The causal effect of the drug for a given subject is the difference between these two potential outcomes. The Average Treatment Effect (ATE) of the drug is the average of the individual-level causal effects.[88]

The fundamental problem of causal inference is that we can see only one of the potential outcomes—depending on whether a subject received the intervention or not—so we can never directly measure the individual treatment effects. In a study that relies on observed behavior (i.e., an observational design), we see the potential outcomes associated with the intervention for subjects that receive the intervention, and the potential outcomes associated with no intervention for those that do not. We could then compare the mean outcomes for the two groups to assess the effect of the intervention. The problem with this comparison is that the average outcome for the nonintervention group can be a misleading estimate of what would have happened to the intervention group in the absence of the intervention. In the blood pressure medication example, the average blood pressure among subjects who do not take the drug could be higher or lower than the average that would have been observed for the intervention group if they had not taken the drug. This difference is defined as *selection bias*: It represents the expected average difference in potential outcomes associated with no intervention between the group that received the intervention and the group that did not.

The sign (positive or negative) and magnitude of selection bias varies from setting to setting and depends on the outcome as well as how the intervention is assigned to (or chosen by) different subjects. For example, in an analysis of the effect of a new medicine on patient death, selection bias will be negative if the medicine is allocated to the sickest patients (because patients who get the medicine are more likely to die, even without the medicine). In contrast, in an analysis of the effect of a training program on earnings, selection bias can be positive if more promising candidates are selected for the program (since they would earn more even without the training).

In an observational study, the simple difference in outcomes between the group that receives the intervention and the group that does not is the sum of the average treatment effect for those who receive the intervention and the selection bias. (See the Appendix for a mathematical statement of this relationship.) In an

87. Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. Educ. Psych. 688 (1974), https://doi.org/10.1037/h0037350.

88. For an explanation in a non-RCT case, see *United States v. Brown*, 299 F. Supp. 3d 976, 999 n.16 (N.D. Ill. 2018) (explaining an ATE in a probability-weighting logistic regression).

RCT, the assignment of the intervention is random, removing any selection bias[89] and ensuring that on average the potential outcomes should be the same for the groups that receive and do not receive the intervention.

This framework provides a slightly different way of thinking about when a multiple regression model, applied to observational data, will yield a coefficient on an indicator variable for being in a group of interest (for example, being over age 40 in an age discrimination case) that can be interpreted causally. In such a setting, the other control variables in the model must fully eliminate any selection bias. In other words, holding constant the observed covariates in the model, all unobserved differences that affect the outcome variable must be the same, on average, for the subjects in the group of interest and those in the rest of the population. This is a very strong condition that needs to be carefully evaluated.

One approach that economists have used to assess the plausibility of the no-selection-bias assumption is to examine how the estimated coefficient on the variable of interest changes as additional control variables are added to the model. In an RCT setting, the estimated coefficient should only change slightly (if at all) as other controls are added to the model, since the key variable of interest is randomly assigned to subjects, and therefore should be uncorrelated with other potential control variables. But in an observational setting, the variable of interest is often correlated with the characteristics of the subjects. In that case, adding control variables may lead to large changes in the estimated coefficient of interest. Altonji et al. suggested that if adding or subtracting observed covariates from the model has little effect on the coefficient of the variable of interest, then it is more plausible that other unobserved factors will have small effects.[90] A finding of relative stability in the coefficient of interest provides confidence that the estimated effect is robust to changes in the specific set of control variables included in the model, and it may offer some assurance of robustness to unobserved factors under the assumptions proposed by Altonji et al.

Another approach that is sometimes feasible is to show that the variable of interest has no effect on a "placebo" outcome: one that is determined by the same factors as the outcome of interest but logically cannot be affected by the variable

89. Note that selection bias refers to the difference in expected values of the outcome between the group that received the intervention and the group that did not in the absence of the intervention. Although a well-designed RCT will have zero selection bias, there could be differences between the characteristics of the two groups that arise randomly—especially if the groups are small. *See* Winston Lin, *Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique,* 7 Annals Applied Stat. 295–318 (2013) (for a discussion of whether these realized differences in characteristics should be accounted for when measuring the treatment effect in an RCT).

90. Joseph G. Altonji, Todd E. Elder & Christopher R. Taber, *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools*, 113 J. Pol. Econ. 151 (2005), https://doi.org/10.1086/426036. Emily Oster presents a related analysis that focuses on how the addition of other controls jointly affects the stability of the coefficient of interest and the explanatory power of the model. Emily Oster, *Unobservable Selection and Coefficient Stability: Theory and Evidence*, 37 J. Bus. & Econ. Stats. 187 (2019), https://doi.org/10.1080/07350015.2016.1227711.

of interest. For example, Rothstein evaluated a regression model that previous researchers had used to show the effect of fourth-grade teacher characteristics on fourth-grade test scores by estimating a similar model but using the same students' *third-grade* test scores as the outcome.[91] He found that fourth-grade teacher characteristics had a positive effect on third-grade scores, suggesting that in this particular setting there was a positive selection bias in the proposed model. Such "falsification" or "placebo" tests—if they are passed—can increase confidence in a causal interpretation of the model.

In cases where an expert is proposing a causal interpretation of the results from a multiple regression model based on observational data, there are two simple questions that need to be addressed:

1. Is there a compelling justification for the assumption (implicit in a causal interpretation) that all omitted factors that could substantially influence the outcome variable are uncorrelated with the variable(s) of interest?
2. What evidence has the expert assembled to support that reasoning?

Pre/Post Design

While randomized controlled experiments offer a very good benchmark, they are rarely conducted in litigation settings, and may not be feasible.[92] Social scientists and legal experts often rely on other non-experimental approaches (sometimes referred to as non-experimental research designs) that can potentially isolate causal effects by trying to mimic the features of an RCT. The simplest of such approaches, widely used in business practice and policy analysis, is a pre/post design (or before/after design). This approach can be applied in situations where data are available on an outcome for subjects that receive an intervention or treatment both *before* and *after* the intervention. A pre/post design interprets the change in the average outcome across subjects as an estimate of the causal effect of the intervention. For example, an analyst may have data on the number of property crimes in a city in the year before and after a new policing policy is introduced. A pre/post estimate of the effect of the policy is just the change in the number of crimes from the year before to the year after the change.

In a Rubin (1974) framework, a pre/post design uses the average outcome in the pre period as an estimate of the mean potential outcome in the absence of treatment. Since the average outcome in the post period is measured for the same subjects, but with treatment, it may appear that a pre/post difference is free of

91. Jesse Rothstein, *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*, 125 Q.J. Econ. 175 (2010), https://doi.org/10.1162/qjec.2010.125.1.175.

92. Although in principle an RCT eliminates selection bias, actual RCTs can have other problems (e.g., missing data) that reintroduce selection bias.

selection bias. By measuring the outcomes in different time periods, however, a new source of bias may be introduced, arising from potential changes over time in what would have happened to the average outcomes of the treated subjects, even if they did not receive the intervention. Validity of a pre/post comparison requires that the average outcome for the group that received the treatment would have remained constant in the absence of treatment (i.e., a constant mean assumption).

Many arguments in law, economics, and politics scholarship revolve around interpretations of pre/post designs and the plausibility of the constant mean assumption. For example, Tracey Meares discusses the debate around the interpretation of crime trends in New York City that were used by the city to defend its "Stop, Question, and Frisk" policies in *Floyd v. City of New York*.[93] A particular concern that arises is the extent to which the mean outcome among the subjects included in the sample being studied has been driven by the timing of police interventions. As another example, Orley Ashenfelter has noted that many people enroll in government training programs after a job loss.[94] Even without training, many of these people would be expected to return to work and experience a rebound in earnings.[95] In such cases, a simple pre/post comparison of earnings for participants in the program will overstate the causal impact of the intervention.

Nevertheless, there are some settings where a pre/post design is compelling, particularly if periods of data are available when there was no change in the treatment. Such data allow an analyst to measure average outcomes for the treated group in the periods before and after the intervention and examine patterns in

93. Tracey L. Meares, *The Law and Social Science of Stop and Frisk*, 10 Ann. Rev. L. & Soc. Sci. 335 (2014), https://doi.org/10.1146/annurev-lawsocsci-102612-134043; Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

94. Orley Ashenfelter, *Estimating the Effect of Training Programs on Earnings*, 60 Rev. Econ. & Stats. 47 (1978), https://doi.org/10.2307/1924332. The phenomenon that an intervention tends to be implemented when the mean outcome is trending downward is so common that it is generally referred to as an *Ashenfelter dip.* Ashenfelter dips have been noted in the timing of turnover of professional sports coaches (e.g., Maria De Paola & Vincenzo Scoppa, *The Effects of Managerial Turnover: Evidence from Coach Dismissals in Italian Soccer Teams*, 13 J. Sports Econ. 152 (2012), https://doi.org/10.1177/1527002511402155) and school principals (e.g., Ashley Miller, *Principal Turnover and Student Achievement,* 36 Econ. Educ. Rev. 60 (2013), https://doi.org/10.1016/j.econedurev.2013.05.004).

95. This is a version of the well-known phenomenon of mean reversion or regression to the mean: the tendency of a statistical process (like individual earnings) to return to its mean value after interruptions that cause unusually high or low values. The concept often arises in securities litigation. *See, e.g.*, IQ Holdings, Inc. v. Am. Com. Lines Inc., No. 6369-VCL, 2013 Del. Ch. LEXIS 234, at *10–12 (Del. Ch. Mar. 18, 2013) (discussing the specific application of mean reversion to calculating the weighted average cost of capital). For another application, in the debate over affirmative action in higher education, see *Students for Fair Admissions, Inc. v. Univ. of N.C.*, 567 F. Supp. 3d 580, 624–25 (M.D.N.C. 2021) (characterizing an expert's averaging as a "regression to the mean"). For a discussion in the context of a pain control study, see *FTC v. QT, Inc.*, 448 F. Supp. 2d 908, 939 (N.D. Ill. 2006).

these averages over time. A simple plot of these means is known as an event study graph.

In cases where an expert is proposing a causal interpretation of the results from a pre/post design (or as unbiased estimates of the effects represented in a theoretical model), a straightforward question needs to be answered: What evidence has the expert assembled to verify the key pre/post design assumption that absent the intervention, the mean outcomes of subjects affected by the intervention would not have changed?

Difference-in-Differences Design

The difference-in-differences (DD) design (sometimes called *diff-in-diff*) extends the pre/post design by adding a comparison group that is not affected by the intervention or treatment.[96] In the simplest version of DD, one group of subjects experiences an intervention or treatment at some point in time, while another group does not. For example, David Card and Alan Krueger considered the changes in employment between February and November 1992 at fast food restaurants in New Jersey, where the state minimum wage was increased in April 1992, relative to the changes over the same period at restaurants in Eastern Pennsylvania, where the state minimum wage remained unchanged.[97] Because the restaurants in Philadelphia and its suburbs were physically close to the restaurants in Camden, New Jersey, and its suburb, it was reasonable to believe that the effects of any other employment-related control covariates would be the same in both locations. In this framework, the DD estimate of the causal effect of the intervention is the change in the average outcome of the treated group, minus the change in the average outcome of the comparison group measured over the same period. This method is widely used in the social sciences and in litigation to examine the causal effects of law changes, firm mergers, and other phenomena.[98]

In the Rubin framework,[99] a DD design uses the average outcome for the treatment group in the pre period, adjusted by the change in outcomes for the

96. In some studies, the untreated group is called a control group, but that nomenclature risks mixing up control groups in randomized trials and comparison groups in difference-in-difference designs, which are not randomly assigned.

97. David Card & Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, 84 Am. Econ. Rev. 772 (1994).

98. For example, Hastings (2004) conducts an analysis of retail gasoline prices at two groups of gas stations in Sothern California: one group of stations within a mile of a station operated by Thrifty, an independent "non-branded" retailer, and another group of stations further away from any Thrifty station. In the middle of her sample period the Thrifty stations were acquired by ARCO, a major petroleum company; she shows that this led to a rise in prices at stations close to the former Thrifty stations.

99. Rubin, *supra* note 87.

comparison group between the pre and post periods, as an estimate of the counterfactual mean potential outcome for the treatment group in the post period (i.e., as an estimate of the mean of the potential outcome in the absence of treatment for the treatment group in the post period). The key assumption is that in the absence of the intervention the mean potential outcomes for the two groups would move in parallel—a counterfactual assumption that is sometimes called a parallel trends assumption. The DD design assumes parallel trends for the two groups, estimates the change in outcomes that would have occurred in the absence of treatment from the average pre/post change for the comparison group, and then subtracts this value from the average pre/post change for the treatment group to derive a causal estimate.

Like the simpler pre/post design, the potential validity of a DD design can be evaluated by using data on outcomes for the treatment and comparison groups in periods before the intervention. Under the parallel trends assumption, the means of the outcomes for the two groups should move in parallel in the pre-intervention periods. A typical event study graph for a DD design plots the mean outcomes of the treatment and comparison groups in periods before and after the intervention relative to the difference that existed in the period just before the intervention, allowing the analyst to directly assess whether the mean outcomes of the two groups moved in parallel in periods prior to the intervention, and whether the gap between the mean outcomes of the treatment and comparison groups widened or narrowed after the date of the intervention.[100]

Generalized Difference-in-Differences Designs

There are many extensions and generalizations of the basic DD design. As with a basic difference-in-differences design, the potential validity of a generalized event-study design can be partly evaluated by plotting the differences in mean outcomes between the treatment and comparison groups for several periods before and after the intervention, removing all differences from the gap in the period immediately before the intervention (i.e., constructing the difference-in-differences between the treatment and comparison groups in each period, relative to the period just before the intervention).[101] In a valid design, all the

100. If it is assumed that the intervention causes a once and for all shift in outcomes, then data for the post-intervention periods should also exhibit parallel trends. For an example relating to gasoline stations, see Hastings, *supra* note 82.

101. For example, the method is "commonly accepted" for monetary damage calculations. Windstream Holdings, Inc. v. Charter Commc'ns Inc. (*In re* Windstream Holdings, Inc.), 627 B.R. 32, 52–53 (Bankr. S.D.N.Y. 2021). *See generally* Messner v. Northshore Univ. HealthSystem, 669 F.3d 802, 818 (7th Cir.), *reh'g denied*, 2012 U.S. App. LEXIS 4778 (7th Cir. Feb. 28, 2012); Smith v. Keurig Green Mt., Inc., 2020 U.S. Dist. LEXIS 172826, at *26–30 (N.D. Cal. Sept. 21, 2020); Lowes Foods LLC v. Burroughs & Chapin Co., 2019 U.S. Dist. LEXIS 100410, at *7

difference-in-differences in the pre-intervention period should be close to zero (apart from sampling error) with no discernible trend.

For example, McCrary studies the effect of class-action lawsuits against discriminatory police-hiring policies on the gap between the fraction of African-American officers in a city's police force and the fraction of African-American residents in the city (a difference he calls the "representation gap").[102] He presents graphs that show the difference-in-differences in the representation gap in cities with a lawsuit relative to the year the lawsuit was filed in that city (i.e., treating the filing date of the lawsuit as the date of the intervention). The comparison group includes data from a relatively large number of other cities that had no lawsuit, as well as from cities with lawsuits in future or past years. McCrary's graphs show that prior to the filing date, the representation gap was negative on average (i.e., the African-American share of police officers was lower than the African-American share of city residents) and was stable or even widening slightly prior to a lawsuit, whereas after the filing date the representation gap narrowed, reflecting increases in the relative hiring of African-American officers.

In cases where an expert is proposing a causal interpretation of the results from a difference-in-differences design (i.e., as unbiased estimates of the causal effects of an intervention), two key questions need to be answered:

1. Has the expert assembled evidence on the validity of the parallel trends assumption—that, in the absence of the intervention, the mean outcomes of the treatment group and the comparison group would move in parallel from period to period? Typically, this evidence will take the form of an analysis of data from several periods before the intervention.
2. If such evidence is not available, what other justification or rationale can be offered for the validity of the parallel trends assumption?

Instrumental Variables Design

An instrumental variables (IV) design isolates a specific part of the variation in a variable of interest across subjects and estimates the causal effect of that part of the variation on an outcome. The method relies on the existence of a so-called instrumental variable that satisfies three critical assumptions: (1) it substantially affects the variable of interest; (2) it is uncorrelated with the unobserved

(D.S.C. Apr. 17, 2019); *In re* Apple Inc. Device Performance Litig., No. 5:18-MD-02827-EJD, 2021 U.S. WL 1022866, (N.D. Cal. Mar. 17, 2021); Ideker Farms, Inc. v. United States, 151 Fed. Cl. 560 (2020).

102. Justin McCrary, *The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police*, 97 Am. Econ. Rev. 318 (2007), https://doi.org/10.1257/aer.97.1.318.

determinants of the outcome; and (3) it has no direct effect on the outcome.[103] For example, Angrist (1990) addresses the question of whether men who served in the military during the Vietnam era had lower or higher earnings later in life as a result of their service.[104] Given that only a small fraction of men of draft-eligible age actually served in the military in this era, he was concerned about selection biases in the observed differences in earnings between veterans and nonveterans. He therefore used the assignment of a relatively low draft lottery number during the draft lotteries of 1970–1972 as an instrumental variable for serving in the military.

Because of the draft process, Angrist noted that men with low lottery numbers had higher rates of military service than other men. As a result, his proposed instrumental variable satisfies condition (1) above. But since lottery numbers were randomly assigned to different birthdays, it is plausible that having a low lottery number was uncorrelated with all other factors that affect earnings (like family background and cognitive ability), and had no direct effect on earnings, satisfying conditions (2) and (3).

IV designs are often used when there is concern that the main variable of interest is partially determined by the same unobserved factors that also affect the outcome—the situation of an endogenous covariate.[105] In the case of the effect of military service on earnings, for example, one might be concerned that latent health conditions affect the probability of military service and have some effect on earnings. The idea of an IV design is to separate out the part of the endogenous covariate that is predicted by the instrumental variable and use only that part to measure the causal effect. In principle, there may be multiple instrumental variables available, though this can lead to complex issues of

103. IV designs appear prominently in federal antitrust litigation. *See* United States v. Aetna Inc., 240 F. Supp. 3d 1, 36 (D.D.C. 2017); *In re* Domestic Drywall Antitrust Litig., 322 F.R.D. 188, 219 (E.D. Pa. 2017); United States v. Am. Express Co., No. 10-CV-4496 (NGG) (RER), 2014 U.S. Dist. LEXIS 87360, at *17–23 (E.D.N.Y. June 24, 2014). See another application in a claim alleging hiring of unauthorized immigrants to depress wages, *Hall v. Thomas*, 753 F. Supp. 2d 1113, 1145–47 (N.D. Ala. 2010).

104. Joshua D. Angrist, *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, 80 Am. Econ. Rev. 313 (1990). In a similar vein, a large number of studies—summarized in Julia Chabrier, Sarah Cohodes & Philip Oreopoulos, *What Can We Learn from Charter School Lotteries?*, 30 J. Econ. Persps. 57 (2016), https://doi.org/10.1257/jep.30.3.57—use the outcomes of admissions lotteries as instrumental variables for attending a charter school in analyzing the effects of charter schools on student test scores.

105. For example, schooling is often modeled as an endogenous regressor in models that relate earnings to schooling, reflecting the belief that some of the same unobserved factors that determine schooling—like ability, ambition, and parental encouragement—also partly determine earnings. David Card, *The Causal Effect of Education on Earnings*, *in* 3 Handbook of Labor Economics (Orley Ashenfelter & David Card eds., 1999).

interpretation and proper inference.[106] This presentation focuses on the simpler case of a single instrumental variable (also known as the just-identified case).

The IV method with a single instrumental variable involves estimating two regression models: hence the term *two-stage least squares* that is widely used to describe the method. The first of these (known as the *first-stage model*) regresses the endogenous variable on the instrumental variable (and any other control variables included in the model). The estimates from this model are then used (in the second stage) to form predicted values of the endogenous regressor for each unit (or subject) in the dataset. Since the prediction is just a weighted average of the instrumental variable and the other control variables, and by assumption the instrumental variable is uncorrelated with the unobserved determinants of the outcome, the associated predictions are also uncorrelated with the unobserved factor. The second-stage model is then a multivariate regression model that relates the outcome to the predicted values of the endogenous covariate and the same control variables included in the first-stage model. The estimated coefficient for the predicted variable is interpreted as a causal estimate of the effect of the endogenous covariate on the outcome.

The key coefficient of the predicted endogenous covariate in the second-stage model can be derived in a different way that illustrates its interpretation. Consider a third model, known as the reduced-form model, that regresses the outcome variable on the instrumental variable and the other controls.[107] The coefficient of the instrumental variable in this model can be interpreted causally because it is assumed to be uncorrelated with the unobserved determinants of the outcome. This coefficient expresses the expected change in the outcome variable for a unit change in the value of the instrumental variable. But since the instrumental variable is assumed to have no *direct* effect on the output, this reduced-form result must arise via the effect of the instrumental variable on the endogenous covariate.

Consider, for example, the analysis of the effect of military service on earnings, using a low lottery number as the instrumental variable. Angrist (in his Table 3) shows that having a low lottery number is associated with about $400 lower earnings per year for men born in 1950 than for men with higher lottery numbers. This is the reduced-form effect of a low lottery number. The entirety of this difference is presumably attributable to the fact that these men were more likely to serve in the military. In fact, Angrist shows that the first-stage model relating military service to lottery numbers implies that they had a

106. John Bound, David A. Jaeger & Regina M. Baker, *Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak*, 90 J. Am. Stat. Ass'n 443 (1995), https://doi.org/10.2307/2291055; James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518 (2002), https://doi.org/10.1198/073500102288618658.

107. The set of variables included in the reduced-form model must be the same as the set included in the first-stage model.

16 percentage-point higher probability of serving in the military than men with higher lottery numbers. Using mathematical notation, suppose that serving in the military has a causal effect of β_1 dollars.. Then the reduced-form effect (–$400) must arise because a 0.16 increase in the probability of military service, multiplied by the effect of military service on earnings (β_1) is –$400. In other words, $-400 = \beta_1 \times 0.16$, implying that $\beta_1 = -400/0.16 = -\$2500$.

In general, the reduced-form effect of the instrumental variable on the outcome is the product of the first-stage effect of the instrumental variable on the endogenous covariate, and the causal effect of a change in the endogenous covariate on the output. This means that one can unscramble the causal effect by dividing the reduced-form coefficient by the first-stage coefficient. (See the Appendix for a mathematical statement of this procedure.) In the case where there is a single instrumental variable, this ratio is exactly equal to the estimated coefficient of the predicted endogenous covariate in the second-stage model.

As another example of the use of an IV design, Dobbie et al. relate defendant outcomes in criminal cases—including case outcomes such as the incidence of guilty pleas and longer-term outcomes like employment rates—to an indicator variable of whether the defendant was detained prior to their trial.[108] Although the authors' dataset includes relatively detailed controls for case characteristics, the authors are concerned that judges set bail based in part on *other* case characteristics that are not in the dataset, and that might also affect later outcomes for the defendant. In essence, pretrial detention is an endogenous covariate.

As an instrumental variable for pretrial detention, the authors use the average rate of pretrial detention in other cases handled by the same bail judge.[109] The plausibility of this choice rests on the fact that in the two large jurisdictions in their sample, bail judges are typically assigned at random to cases (conditional on time of day and day of the week of the hearing). The authors show that judges with a higher rate of pretrial detention in other cases are more likely to set harsher bail terms that lead to a higher probability of detention, satisfying condition (1) above. They also show that the instrumental variable is uncorrelated with a long list of characteristics that are highly predictive of pretrial detention (such as measures of the defendant's previous crime record). This lack of correlation does not *prove* that conditions (2) and (3) are satisfied but is consistent with their assertion that judges are as good as randomly assigned and would fail if harsher bail judges were assigned to specific types of cases. The authors' analysis of the judge assignment process

108. Dobbie et al., *supra* note 83.

109. This is known as the leave-out mean. To account for the fact that certain kinds of cases are more likely at certain times or days of the week, Dobbie et al. use a leave-out mean of a regression-adjusted pretrial detention rate, where the adjustment factors include time-of-day and day-of-week indicators. *Id*.

makes a plausible case that conditions (2) and (3) are satisfied.[110] Of course, not all assignments are random, in part because not all judges are on the same selection "wheel." However, empirical evidence supports the view that for broad categories of cases, the judicial assignment process appears to be random.[111]

As in the Rubin framework, each subject has two potential outcomes, representing the value of their outcome if they receive the intervention or not.[112] Among all the subjects assigned the high value of the instrumental variable, there is a fraction of compliers whose observed outcome is their potential outcome with treatment. Among the subjects assigned the low value of the instrumental variable are the same fraction of compliers whose observed outcomes are their potential outcomes without treatment. The share of compliers in the two groups, and their distributions of potential outcomes, is the same in both cases because the instrumental variable is working as if the individuals were randomly assigned. As a result, the expected difference in average outcomes between subjects that are assigned high and low values of the instrumental variable is the difference in potential outcomes among the compliers, multiplied by their share in the sample. This product is estimated by the coefficient of the instrumental variable in the reduced-form equation of the IV procedure. Dividing this coefficient by the first-stage coefficient of the instrumental variable (which, as noted earlier, is the IV estimate of the causal effect) yields an estimate of the average treatment effect among the compliers. In many applications the complier group can be relatively small: For example, in Angrist's study of the Vietnam-era draft lottery, the group was composed of only about 15% of draft-age men.[113] Thus the estimated effect derived from an IV design does not necessarily provide an estimate of the average treatment effect for the entire subject population.

110. For an extension of Rubin's potential-outcomes framework to an IV design in which the endogenous covariate is an indicator for receiving an intervention (e.g., serve in the military or not), and the instrumental variable is also dichotomous (e.g., receive a high or low draft lottery number) that is distributed randomly or (by assumption) *as if* randomly, see Guido W. Imbens & Joshua D. Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 Econometrica 467 (1994), https://doi.org/10.2307/2951620. For example, an instrumental variable based on the last four digits of a Social Security number is not randomly assigned but may be "as good as" randomly assigned. Often it is assumed that a variable is as good as randomly assigned conditional on a set of basic controls. For example, whether a person wins a lottery is random holding constant the number of tickets they held, but not unconditionally.

111. Orley Ashenfelter, Theodore Eisenberg & Stewart J. Schwab, *Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes,* 24 J. Legal Stud. 257 (1995), https://doi.org/10.1086/467960.

112. The never takers and always takers also have two potential outcomes, but their outcomes are not affected by which value of the instrumental variable they are assigned, because the instrumental variable does not affect their participation in the intervention, and by assumption changes in the instrumental variable do not directly affect outcomes.

113. Angrist, *supra* note 104.

In applications of IV it is important to assess the validity of the three assumptions noted above. Violations of these assumptions can lead to large biases in the estimated causal effects obtained from an IV design—in some cases, even larger than the bias associated with a simple multivariate regression model that ignores concerns about the endogenous variable (i.e., the purely observational design, estimated by ordinary least squares (OLS)). With that in mind, an IV estimate should always be compared directly with the associated OLS estimate. If the two estimates are substantively different, there should also be a careful discussion of the sources of bias in the OLS estimate.

The first assumption for an IV design is that the instrumental variable has a substantial effect on the variable of interest. Since this effect is measured by the coefficient of the instrumental variable in the first-stage model, standard practice is to present the first-stage model and evaluate the sign (positive or negative), size, and statistical significance of this coefficient. The sign is important because the logic of an IV design rests on the idea that changes in the instrumental variable cause systematic changes in the endogenous covariate: the direction of this causal effect is reflected in the sign of the first-stage coefficient and must be interpretable.[114] Size and significance are important because the IV estimate of the causal effect is based on the response of the outcome to the part of the endogenous covariate that is attributable to variation in the instrumental variable. When the first-stage coefficient is large in magnitude and highly statistically significant, this part of the variation in the endogenous covariate can be more reliably identified.[115] As an example of how sign, size, and significance can be assessed visually, Dobbie et al. plot the mean of their endogenous covariate for various ranges of their instrumental variable and show that there is a strong relationship of the expected sign.[116] In addition, they report the estimated coefficient of the instrumental variable in their first-stage equation.

The second assumption is that the instrumental variable is uncorrelated with all unobserved determinants of the outcome. This condition will be satisfied if

114. Isaiah Andrews and Timothy B. Armstrong note that use of information on the sign of the first-stage coefficient can substantially improve the performance of IV estimators. Isaiah Andrews & Timothy B. Armstrong, *Unbiased Instrumental Variables Estimation Under Known First-Stage Sign*, 8 Quantitative Economics, 479 (2017), https://doi.org/10.3982/QE700.

115. In the case of multiple instrumental variables, a conventional "rule of thumb" is that the F-statistic that provides a means of testing the goodness of fit of the instrumental variables in the first-stage equation has to exceed ten. James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518 (2002), https://doi.org/10.1198/073500102288618658. With a single instrumental variable, such a cutoff is not necessarily best practice. Instead, it is recommended that analysts report so-called Anderson-Rubin confidence intervals for the estimated causal effect. *See* Isaiah Andrews, James H. Stock & Liyang Sun, *Weak Instruments in IV Regression: Theory and Practice*, 11 Ann. Rev. Econ. 727 (2019), https://doi.org/10.1146/annurev-economics-080218-025643. These confidence intervals incorporate uncertainty from the first-stage model.

116. Dobbie et al., *supra* note 83.

the instrumental variable is randomly assigned. Otherwise, the burden of persuasion should be on the analyst to present evidence that the instrumental variable is as good as randomly assigned, holding constant a basic set of control variables in the model. One approach (used by Dobbie et al.) is to show that the instrumental variable is uncorrelated with predetermined subject characteristics that are predictive of the outcome (at least once the main control variables are introduced).[117] Another complementary approach is to show that the reduced-form effect of the instrumental variable on the outcome remains relatively stable as additional control variables are added to the model (holding constant the set of basic controls needed to ensure the instrumental variable is as good as randomly assigned). This is an adaptation of the argument of Altonji et al. for observational designs previously discussed.[118] But since the reduced-form regression is itself just a multiple regression model with one key covariate of interest (the instrumental variable), the same arguments apply.

The third assumption is that the instrumental variable does not itself directly affect the outcome—sometimes called an exclusion restriction. There are two related concerns over this assumption in particular settings. First, sometimes an instrumental variable may in fact have its own causal effect.[119] This is not normally an issue if the instrumental variable is based on a randomizing process, as in the draft-lottery study of Angrist,[120] or a quasi-randomizing process, as in the study based on judge assignments by Dobbie et al.[121] In other settings, however, the exclusion restriction must be justified on theoretical grounds. For example, in models of consumer demand for differentiated products that are widely used in antitrust analyses, some analysts argue that sums of characteristics of competing products, interacted with whether the competing products are supplied by the same firm or another firm, are plausible instrumental variables for product prices.

Second, even if the instrumental variable is as good as randomly assigned, there may be another endogenous determinant of the outcome that is also affected by the instrumental variable. In this case, the IV design confounds the effects of the two endogenous variables—a so-called multiple channel problem—possibly overstating or understating the causal effect of the variable of interest. For

117. Such an analysis is conducted by Chan et al., who show that medical patient characteristics help predict the probability of a pneumonia diagnosis (their outcome variable) but do not predict the instrumental variable in their model once they control for patient age, a variable they argue is needed to ensure that the instrumental variable's distribution is as good as random across patients. David Chan, Matthew Gentzkow & Chuan Yu, *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*, 137 Q.J. Econ. 729 (2022), https://doi.org/10.1093/qje/qjab048.

118. Altonji et al., *supra* note 90.

119. For example, some studies of the effect of education on earnings have used parental education as an instrumental variable. *See* Card, *supra* note 105.

120. Angrist, *supra* note 104.

121. Dobbie et al., *supra* note 83.

example, an instrumental variable that increases the education level of young people of a given age also typically reduces the years of work experience they have completed since finishing their schooling (since someone of a given age with more education has had less time to work). Since both education and work experience affect earnings, it may be inappropriate to attribute the difference in earnings between people with different values of the instrumental variable entirely to their differences in schooling. Arguments ruling out such multiple-channel concerns are typically based on theoretical grounds.

In cases where an expert is proposing a causal interpretation of the results from an instrumental variables design (or as unbiased estimates of the effects represented in a theoretical model), a series of key questions must be answered:

1. What is the instrumental variable (or set of instrumental variables)? What is the reasoning underlying the use of this instrumental variable?
2. What is the sign (positive or negative) and size of the coefficient representing the effect of the instrumental variable in the first-stage model? The method depends critically on the assumption that the first-stage model is describing the effect of the instrumental variable on the variable of interest. Therefore, the first-stage equation must be interpretable; it must make sense in light of the claim at issue.
3. Is the measured effect of the instrumental variable in the first-stage model statistically significant?
4. What is the reasoning underlying the implicit assumption—necessary for a causal interpretation of estimates from an instrumental variables design—that the instrumental variable is uncorrelated with any unobserved determinants of the outcome variable in the second-stage model?
5. What evidence has the expert assembled to support that reasoning?
6. What is the reasoning underlying the implicit assumption—necessary for a causal interpretation of estimates from an instrumental variables design—that the instrumental variable has no direct effect on the outcome variable?

More Advanced Regression Methods

Fixed-Effects and Random-Effects Regression Models

As the volume of data that is available has grown over time, social scientists have often utilized both time-series and cross-section regression models that allow for the possibility that differences in the values of the dependent variable can be captured in differences in the constant term in the regression. A fixed-effects regression model will include a series of dummy variables as well as the usual set of

relevant covariates.[122] In a cross-section regression model, the fixed effects might account for variation in certain characteristics of individuals in the sample, such as geographic location.[123] In a time-series model that includes daily or weekly data, the fixed-effects model might account for certain time periods, such as year or month.

There are benefits and costs associated with the inclusion of a group of fixed-effects variables in a regression model. The benefit is that the fixed-effects model can reduce or eliminate the possibility of omitted variable bias. Suppose, for example, that in a study of the effect of a merger on the prices paid by consumers for different models and brands of clothes dryers, one believes it appropriate to include a set of covariates that account for the brand, the size of the machine, the number of cycles, and whether the dryer is electric or gas.[124] As an alternative, however, the inclusion of a series of fixed-effects dummy variables representing each individual brand and model of clothes dryer might provide a richer analysis.

The cost is that the inclusion of the fixed-effects variables might mask a more complete characterization of the phenomenon being studied. To continue the clothes dryer example, suppose that companies frequently introduce (and withdraw) new models, with each model typically only lasting in the market a year or less. In this case, controlling for individual model effects may preclude a complete analysis of the effect of the merger on prices, since only models that were in the market both before and after the merger contribute to the estimation of the effect of the merger on prices. A more instructive regression model would have a more limited set of controls that allow pre- and post-merger comparisons to be made across all the models offered in the market in each month, rather than just the subset of brands that were present in the market before and after the merger.

The fixed-effects model offers a reasonable approach when we believe that the differences in the dependent variable measured across units of observation can be viewed as vertical shifts in the regression line, and when the basic units of interest (in the example, models of clothes dryers) are observed across all time periods. But if the classification of basic units varies over time (as in the dryer example), it may be more appropriate to consider a model with a set of covariates that are observed consistently, and to assume that the variation in the outcome variable at the group level is an extra error component that is randomly

122. *See, e.g.*, Conrad v. Jimmy John's Franchise, LLC, No. 18-CV-00133, 2021 U.S. Dist. LEXIS 33933 (N.D. Ill. Jan. 10, 2022); United States v. Mariner Health Care, 552 F. Supp. 3d 938 (N.D. Cal. 2021); Sidibe v. Sutter Health, No. 12-CV-04854, 2020 U.S. Dist. LEXIS 136657 (N.D. Cal. July 30, 2020); Nexteel Co. v. United States, 569 F. Supp. 3d 1354 (Ct. Int'l Trade 2022); *In re* Allstate Corp. Sec. Litig., No. 16-C-10510, 2022 U.S. Dist. LEXIS 52616 (N.D. Ill. 2022).

123. *See, e.g.*, Annex Books, Inc. v. City of Indianapolis, 581 F.3d 460 (7th Cir. 2009).

124. *See* Orley C. Ashenfelter, Daniel S. Hosken & Matthew C. Weinberg, *The Price Effects of a Large Merger of Manufacturers: A Case Study of Maytag-Whirlpool,* 5 Econ. J.: Econ. Pol'y 239 (2013), https://doi.org/10.1257/pol.5.1.239.

distributed across the individual units of observation.[125] In essence, using a random-effects model is more appropriate when the goal of a study is to understand behavior in the underlying population of all possible cross-sectional units, whereas the fixed-effects model is focused on an understanding of the behavior of units within the groups that are identified by the fixed effects.

Methods for Estimating Regression Models

The estimated parameters of the basic linear multiple regression model (i.e., the regression coefficients) are most easily and most often estimated using computer programs that seek to minimize the sum of the squares of the regression residuals. When the errors in the regression model fit a normal distribution, the parameter estimates will lie within a normal distribution whose mean is an unbiased estimate of the population mean and whose variance is the minimum variance among all possible estimators. When certain assumptions of the model are violated, the least-squares method can still be highly effective. For example, when heteroscedasticity is an issue (the error variance itself is known to be variable), then weighted least squares can be appropriate. In this case, the data might be weighted by the inverse of an estimate of the standard deviation of the error term.

When the errors are believed to be far from normally distributed, maximum likelihood estimation offers a more robust methodology than least squares. For example, probit and logit models that are widely used in studies of binary outcome variables are fit by this method, as are models of censored outcomes (such as the length of time that a patient survives after a medical procedure, which will be censored if the patient is still alive at the time the data are collected). In essence, maximum likelihood estimation determines values for the parameters of the regression model that maximize the likelihood that the process described by the model produced the data that were observed. The methodology provides a flexible approach that is suitable for a wide variety of models, including models for which the basic regression framework is unsuitable.[126] It has the advantage that, for large samples, the resulting estimates will be unbiased, and if

125. *See, e.g.*, Newman v. McNeil Consumer Healthcare, No. 10-C-1541, 2013 U.S. Dist. LEXIS 113438 (N.D. Ill. Mar. 29, 2013); *In re* Lipitor (Atorvastatin Calcium) Mktg. Sales Pracs. & Prods. Liab. Litig., 145 F. Supp. 3d 573 (D.S.C. 2015).

126. This approach is often used in voting rights cases. *See* Robinson v. Ardoin, 605 F. Supp. 3d 759 (M.D. La. 2022) (describing the ecological inference (EI) model utilized by an expert as using "maximum likelihood statistics to produce estimate of voting patterns by race") (citation omitted); Fabela v. City of Farmers Branch, No. 3:10-CV-1425-D, 2012 U.S. Dist. LEXIS 108086, at *38 n.22 (N.D. Tex. Aug. 2, 2012) (likewise detailing that the benefit of the EI model in analyzing election data is that it does not rely "on an assumption of linearity but instead uses a maximum likelihood estimation" that "provides accurate confidence intervals").

the model is correctly specified, those estimates will have the smallest possible variance.

Why not use maximum likelihood estimation in all cases? In a basic regression model, if one assumes that the errors are normally distributed, then a standard regression approach yields estimated coefficients that are the same as the maximum likelihood estimates. Thus, in many applications, the two approaches are the same. More generally, a standard-regression approach has the advantage that the effects of problems like omitted variables, endogeneity, and measurement error in the explanatory variables are well understood, and often can be assessed in ways that have been extensively developed in the applied econometrics literature.

Measuring the goodness of fit in a model that is estimated using a maximum likelihood method can be complicated, especially when the model is inherently nonlinear and when there is a complex error structure. In simple probit and logit models, measures of goodness of fit known as pseudo *R*-squared measures are widely used.[127] Different versions of models fit by maximum likelihood (say, with more or fewer covariates) can also be compared by a likelihood-ratio test, which measures the percentage improvement in the likelihood of the more complex model over the simpler model.[128]

The Expert

Multiple regression and other forms of complex statistical models are taught to students in extremely diverse fields, including but not limited to statistics, economics, political science, sociology, psychology, anthropology, public health, and history. Nonetheless, the methodology is difficult to master, necessitating a combination of technical skills (the science) and experience (the art). This naturally raises two questions:

1. Who should be qualified as an expert?
2. When and how should the court appoint an expert to assist in the evaluation of issues relating to multiple regression?

127. One version of pseudo *R*-squared was developed in Daniel McFadden, *Conditional Logit Analysis of Qualitative Choice Behavior, in* Frontiers in Econometrics 105 (Paul Zarembka ed., 1973).

128. The significance of the improvement in the model can be evaluated using the chi-square distribution. The chi-square distribution is a standard reference distribution in statistics that represents the sum of the squares of a group of independent standardized normal random variables.

Who Should Be Qualified as an Expert?

Any individual with substantial training in and experience with multiple regression and other statistical methods may be qualified as an expert. A doctoral degree in a discipline that teaches theoretical or applied statistics, such as economics, history, and psychology, usually signifies to other scientists that the proposed expert meets this preliminary test of the qualification process. It is noteworthy that a proposed expert whose only statistical tool is regression analysis may not be able to judge when a statistical analysis should be based on an approach other than regression analysis.[129]

The decision to qualify an expert in regression analysis rests with the court. Clearly, the proposed expert should be able to demonstrate an understanding of the discipline. Publications relating to regression analysis in peer-reviewed journals, active memberships in related professional organizations, courses taught on regression methods, and practical experience with regression analysis can indicate a professional's expertise. However, the expert's background and experience with the specific issues and tools that are applicable to a particular case should also be considered during the qualification process. Thus, if the regression methods are being utilized to evaluate damages in an antitrust case, the qualified expert should have sufficient qualifications in economic analysis as well as statistics. An individual whose expertise lies solely with statistics will have a limited ability to evaluate the usefulness of alternative economic models. Similarly, if a case involves eyewitness identification, a background in psychology as well as statistics may provide essential qualifying elements.

Should the Court Appoint a Neutral Expert?

There are conflicting views on the issue of whether court-appointed experts should be used. In complex cases in which two experts are presenting conflicting statistical evidence, the use of a "neutral" court-appointed expert can be advantageous, although courts would need to find the funds to appoint such an expert. There are those who believe, however, that there is no such thing as a truly neutral expert. In any event, if an expert is chosen, that individual should have substantial expertise and experience—ideally, the expert should be someone who is respected (and trusted) by both plaintiffs and defendants.[130]

129. To illustrate, a case involving allegations of the pass-through of a manufacturers' price-fixing agreement may require testimony by experts with statistical as well as economics training.

130. Judge Posner notes in *In re High Fructose Corn Syrup Antitrust Litig.*, 295 F.2d 651, 665 (7th Cir. 2002), "the judge and jury can repose a degree of confidence in his testimony that it could not repose in that of a party's witness. The judge and the jury may not understand the neutral expert perfectly but at least they will know that he has no axe to grind, and so, to a degree anyway, they will

The appointment of such an expert is likely to influence the presentation of the statistical evidence by the experts for the parties in the litigation. The neutral expert will have an incentive to present a balanced position that relies on broad principles for which there is consensus amongst the community of experts. As a result, the parties' experts can be expected to present testimony that confronts core issues that are likely to be of concern to the court and testimony that is sufficiently balanced to be persuasive to the court-appointed expert.[131]

Rule 706 of the Federal Rules of Evidence governs the selection and instruction of court-appointed experts. In particular,

1. The expert should be notified of the duties through a written court order or at a conference with the parties.
2. The expert should inform the parties of findings orally or in writing.
3. If deemed appropriate by the court, the expert should be available to testify and may be deposed or cross-examined by any party.
4. The court must determine the expert's compensation.[132]
5. The parties should be free to utilize their own experts.

Although not required by Rule 706, it will usually be advantageous for the court to opt for the appointment of a neutral expert as early in the litigation process as possible. It will also be advantageous to minimize any ex parte contact with the neutral expert; this will diminish the possibility that one or both parties will come to the view that the court's ultimate opinion was unreasonably influenced by the neutral expert.

Rule 706 does not offer specifics as to the process of appointment of a court-appointed expert. One possibility is to have the parties offer a short list of possible appointees. If there was no common choice, the court could select from the combined list, perhaps after allowing each party to exercise one or more peremptory challenges. Another possibility is to obtain a list of recommended experts from a selection of individuals known to be experts in the field.

be able to take his testimony on faith." Such an appointment is not an obligation of Federal Rule of Evidence 706. *See* Stevenson v. Windmoeller & Hoelscher Corp., 39 F.4th 466 (7th Cir. 2022).

131. For a discussion of the presentation of expert evidence generally, including the use of court-appointed experts, see Samuel R. Gross, *Expert Evidence*, 1991 Wis. L. Rev. 1113 (1991). Some critique court-appointed experts as biased for the prosecution. *See* J. Blais & A.E. Forth, *Prosecution-Retained Versus Court-Appointed Experts: Comparing and Contrasting Risk Assessment Reports in Preventative Detention Hearings*, 38 L. & Hum. Behav., 531 (2014), https://doi.org/10.1037/lhb0000082. For one study evaluating the newer practice of "hot tubbing," or concurrent expert testimony, see Jennifer T. Perillo et al., *Testing the Waters: An Investigation of the Impact of Hot Tubbing on Experts from Referral Through Testimony*, 45 L. & Hum. Behav. 229 (2021), https://doi.org/10.1037/lhb0000446.

132. Although Rule 706 states that the compensation must come from public funds, complex litigation may be sufficiently costly as to require that the parties share the costs of the neutral expert.

Presentation of Statistical Evidence

The costs of evaluating statistical evidence can be reduced and the precision of that evidence increased if the discovery process is used effectively. In evaluating the admissibility of statistical evidence, courts should consider the following issues:

1. Has the expert provided sufficient information to replicate the multiple regression analysis?
2. Are the expert's methodological choices reasonable or are they arbitrary and unjustified?
3. What disagreements exist regarding the data on which the analysis is based?

In general, a clear and comprehensive statement of the underlying research methodology is a requisite part of the discovery process. The expert should be required to reveal both the nature of the experimentation carried out and the sensitivity of the results to the data and to the methodology.

The following suggestions can substantially improve the discovery process:

1. To the extent possible, the parties should be encouraged to agree to use a common database. Even if disagreement about the significance of the data remains, early agreement on a common database can help focus the discovery process on the important issues in the case.
2. A party that offers data to be used in statistical work, including multiple regression analysis, should be encouraged to provide the following to the other parties: (a) a detailed record of the underlying sources of the data; (b) a data file (ideally, in a commonly used format such as comma-separated values format); (c) complete documentation of the variables in the file; (d) computer programs that were used to generate the data; and (e) explanatory notes associated with the computer programs. The documentation should be sufficiently complete and clear so that the opposing expert can reproduce all the statistical work.
3. A party offering data should make available the personnel involved in the compilation of such data to answer the other party's technical questions concerning the data and the methods of collection or compilation.
4. A party proposing to offer an expert's regression analysis at trial should ask the expert to fully disclose, along with the database and its sources,[133] (a) the method of collecting the data and (b) the methods of analysis. When possible, this disclosure should be made sufficiently in advance of trial so that the opposing party can consult its experts and prepare

133. These sources would include all variables used in the statistical analyses conducted by the expert, not simply those variables used in a final analysis on which the expert expects to rely.

cross-examination. The court must decide on a case-by-case basis where to draw the disclosure line.

These suggestions are motivated by the objective of improving the discovery process to make it more informative. The fact that these questions may raise some doubts or concerns about a particular regression model should not be taken to mean that the model does not provide useful information. It does, however, take considerable skill for an expert to determine the extent to which information is useful when the model being utilized has some shortcomings.

Which Database Information and Analytical Procedures Will Aid in Resolving Disputes Over Statistical Studies?

To help resolve disputes over statistical studies, an expert should follow the guidelines below when presenting database information and analytical procedures.[134]

The expert should

1. Clearly state the objectives of the study, as well as the time frame to which it applies and the statistical population to which the results are being projected.
2. Report the units of observation (e.g., consumers, businesses, or employees).
3. Clearly define each variable.
4. Clearly identify the sample for which data are being studied,[135] as well as the method by which the sample was obtained.
5. Reveal if there are missing data, whether caused by a lack of availability (e.g., in business data) or nonresponse (e.g., in survey data), and the method used to handle the missing data (e.g., deletion of observations).
6. Report investigations into errors associated with the choice of variables and assumptions underlying the regression model.
7. If samples were chosen randomly from a population (i.e., probability sampling procedures were used),[136] make a good-faith effort to provide an estimate of a sampling error, the measure of the difference between the

134. For a more complete discussion of these requirements, see *The Evolving Role of Statistical Assessments as Evidence in the Courts* 256 (Stephen E. Fienberg ed., 1989) (Recommended Standards on Disclosure of Procedures Used for Statistical Studies to Collect Data Submitted in Evidence in Legal Cases).

135. The sample information is important because it allows the expert to make inferences about the underlying population.

136. In probability sampling, each representative of the population has a known probability of being in the sample. Probability sampling is ideal because it is highly structured, and in principle

sample estimate of a parameter (such as the mean of a dependent variable under study), and the (unknown) population parameter (the population mean of the variable).[137]

8. Report the set of procedures that were used to minimize sampling errors if probability sampling procedures were not used.

Appendix: The Basics of Multiple Regression

Introduction

This appendix illustrates, through examples, the basics of multiple regression analysis in legal proceedings. Often, visual displays are used to describe the relationship between variables that are used in multiple regression analysis. Figure 2 is a scatterplot that relates scores on a job aptitude test (shown on the *x*-axis) and job performance ratings (shown on the *y*-axis). Each point on the scatterplot shows what a particular individual scored on the job aptitude test and how that individual's job performance was rated. For example, the individual represented by Point A in Figure 2 scored 49 on the job aptitude test and had a job performance rating of 62.

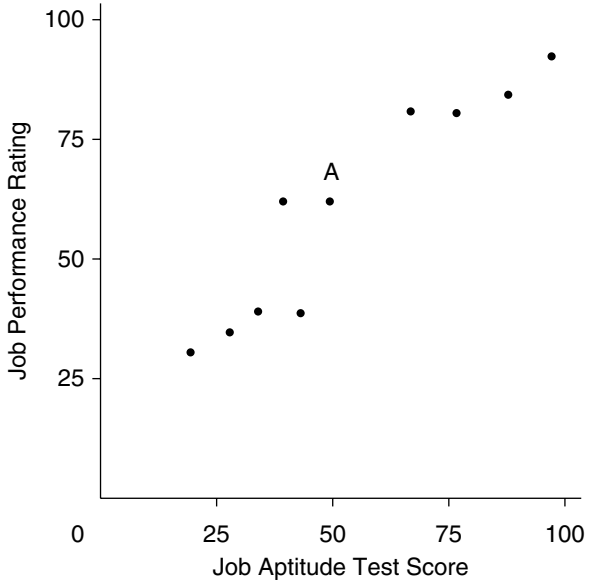
The degree of linear association between two variables can be summarized by a correlation coefficient, which ranges in value from –1 (a perfect linear negative relationship) to +1 (a perfect linear positive relationship).[138] Figure 3 depicts three possible relationships between the job-aptitude variable and the job-performance variable. In Figure 3(a), there is a positive correlation: In general, higher job performance ratings are associated with higher aptitude test scores, and lower job performance ratings are associated with lower aptitude test scores. In Figure 3(b), the correlation is negative, meaning that higher job performance ratings are associated with lower aptitude test scores, and lower job performance ratings are associated with higher aptitude test scores. Positive and negative correlations can

it can be replicated by others. Nonprobability sampling is less desirable because it is often subjective, relying to a large extent on the judgment of the expert.

137. Sampling error is often reported in terms of standard errors or confidence intervals. See Appendix, below, for details.

138. Loosely speaking, the correlation coefficient summarizes the extent to which a scatter plot of variable *Y* against variable *X* lies on a line. In fact, the *R*-squared of a simple linear regression model relating variable *Y* to variable *X* and a constant is the square of the correlation coefficient (which is the reason for the name *R*-squared). In cases where *Y* and *X* are related, but in a nonlinear way, the correlation coefficient is a less useful summary. For example, if *Y* is perfectly determined as a quadratic function of *X*, and *X* takes on values of -2, -1, 0, 1, and 2, then the correlation coefficient relating to *Y* and *X* is precisely 0.

Figure 2. Scatterplot of scores on a job aptitude test relative to job performance rating.



be relatively strong or relatively weak. If the relationship is sufficiently weak, there is effectively no correlation, as is illustrated in Figure 3(c).

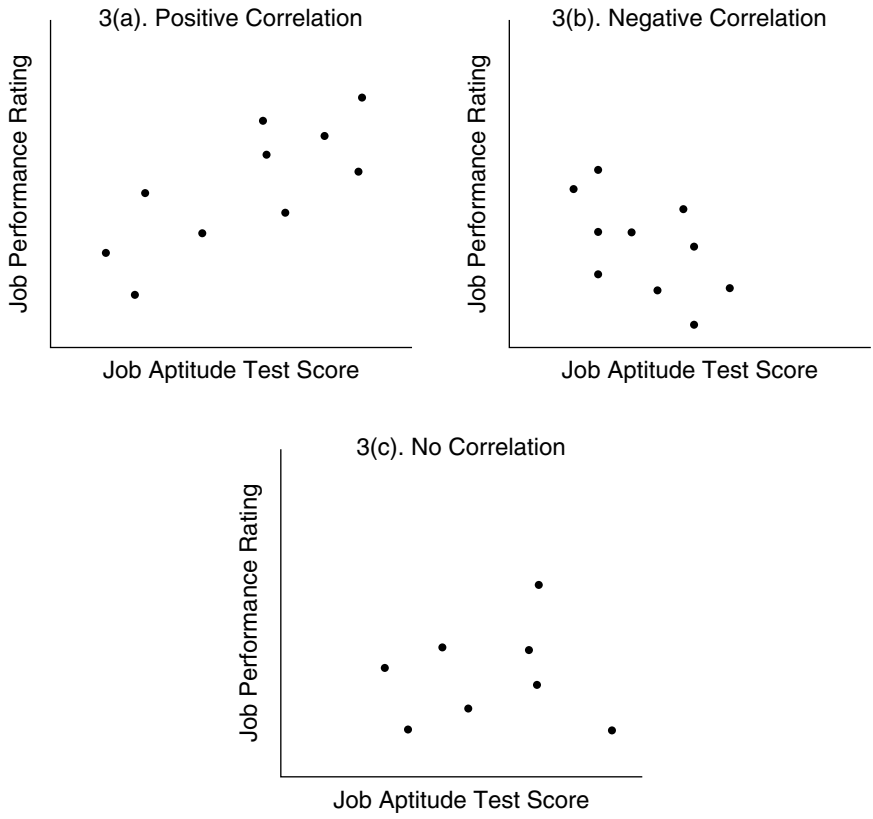
Multiple regression analysis goes beyond simple correlation by deriving the "best fitting" linear function of all the covariates in predicting the dependent variable. For example, if average job performance ratings depend on aptitude test scores, age, and education, multiple regression analysis can use information about the joint behavior of job performance and the three explanatory variables in the observed sample to derive the best fitting linear function of aptitude test scores, age, and education in predicting job performance.

Many aspects of regression analysis can be illustrated in the case where there is only a single explanatory variable. A linear relationship between an explanatory variable X and a dependent variable Y is just the equation of a line:

$$Y = a + bX \tag{1}$$

In this equation, a is the intercept of the line (i.e., the height of the line when it intersects the y-axis with $X = 0$), and b is the slope (the change in the dependent variable associated with a 1-unit change in the explanatory variable). It is important to note the "1-unit change" interpretation, because sometimes the units of measurement change (e.g., from ounces to pounds), and in that case the

Figure 3. Correlation between the job aptitude variable and the job performance variable: (a) positive correlation, (b) negative correlation, (c) weak relationship with no correlation.



meaning of a "1-unit change" also changes. For example, if b is the slope when X is measured in pounds, then the slope when X is measured in ounces will be $b/16$, since each ounce is only 1/16 of a pound.

Unless Y is perfectly linearly related to X, equation (1) does not fully describe the determination of Y. Instead, there is a third term in the equation representing the part of Y that is not attributable to the linear effect of X:

$$Y = a + bX + \varepsilon \tag{2}$$

The term ε, usually called the residual term or the error term, incorporates all the determinants of Y apart from the intercept (also called a *constant term*) and

the linear effect of X. In many cases, an analyst will assume that the average value of the residual term is 0 for each value of X. In that case, the fact that there is a high or low value of X has no bearing on the likely sign (positive or negative) or magnitude of the other unobserved determinants of Y. As discussed above, however, in the interpretation of regression models, one of the critical issues is the plausibility of this assumption.

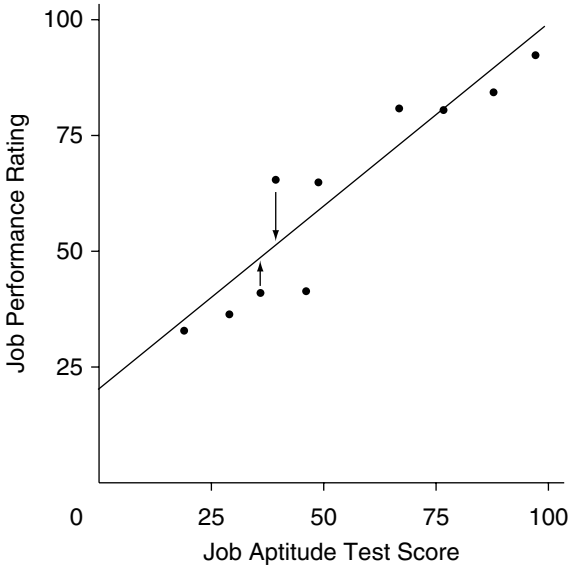
A linear regression model typically is estimated using the method of least squares (also known as ordinary least squares, or OLS), in which the values of the coefficients a and b are calculated so that the sum of the squared residual values (measured by the arrows shown in Figure 4) is made as small as possible. Note that by squaring the residuals, positive and negative deviations of the value of Y from its predicted value ($a + bX$) are given equal weight. Minimization of the squared residuals also means large deviations of Y from its predicted value receive substantially more weight in the determination of a and b than small deviations.

Figure 4, for example, shows the fitted regression line relating aptitude scores (the X variable) to job performance (the Y variable). When the aptitude test score is 0, the predicted (average) value of the job performance rating is the intercept, 18.4. Also, for each additional point on the test score, the job performance rating increases 0.73 units, which is given by the slope 0.73. Thus, the estimated regression line is[139]

$$\hat{Y} = 18.4 + 0.73X$$

139. In discussions of estimated regression models, it is standard to denote the predicted value with a carrot (or "hat") symbol.

Figure 4. Regression line.



Multiple Regression Model

When there are an arbitrary number of explanatory variables, the linear regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{3}$$

where Y represents the dependent variable, such as the salary of an employee, and $X_1 \ldots X_k$ represent the explanatory variables (e.g., education, years of work experience, location in a major metropolitan area, etc.). As in equation (2) above, the error term ε represents the collective influence of all omitted factors influencing Y. Sometimes an analyst will assume that these factors are purely random and can be represented as independent observations drawn from a normal distribution with mean 0 and some standard deviation. Other times, an analyst will leave the precise distribution of ε unspecified.

In a multiple regression, each of the explanatory variables has its own coefficient (e.g., β_1 for the first variable, β_2 for the second).[140] In addition, there is a constant or intercept term, β_0, which has the interpretation of the expected value (or mean) of Y when all of the explanatory variables have a value of 0 (similar to the interpretation of the constant term a in equation (2) above). The full set of coefficients (also referred to as the parameters of the regression model) is usually estimated by ordinary least squares, which again selects the values to minimize the sum of the squared residuals.

Each coefficient β_k measures how the dependent variable Y responds, on average, to a change in the corresponding covariate X_k, holding constant the effects of the other covariates. This can be seen mathematically from equation (3): When all other covariates are held constant, and the value of the residual term is also held constant, a 1-unit increase in the k^{th} covariate will change the value of the dependent variable by an amount β_k.

An important feature of the use of ordinary least squares to estimate the coefficients in equation (3) is that one can give a precise interpretation of how the estimated coefficients are obtained in such a way as to reflect the "holding constant" of the other covariates. Consider the following three-step procedure. First, calculate the residuals from a regression of Y on all covariates other than X_k. These residuals reflect the part of Y that cannot be explained by the other control variables. Second, calculate the residuals of a regression of X_k on all the other covariates. These residuals reflect the part of X_k that cannot be explained by the other control variables. Third and finally, regress the first residual variable on the second residual variable. The resulting coefficient will be identical to β_k. Thus the coefficient in a multiple regression represents the slope of the line "Y, adjusted for all covariates other than X_k versus X_k adjusted for all the other covariates."[141]

Most statisticians and applied economists use the least-squares regression technique because of its simplicity and its desirable statistical properties. As a result, it is also used frequently in legal proceedings.

140. The variables themselves can appear in many different forms. For example, Y might represent the logarithm of an employee's salary, and X_1 might represent the logarithm of the employee's years of experience. The logarithmic representation is appropriate when Y increases exponentially as X increases—for each unit increase in X, the corresponding increase in Y becomes larger and larger. For example, if an expert were to graph the growth of the U.S. population (Y) over time (t), the following equation might be appropriate: $\log(Y) = \beta_0 + \beta_1 \log(t)$.

141. In econometrics, this is known as the Frisch-Waugh-Lovell theorem. Note that if X_k can be perfectly predicted by the other explanatory variables, then there will be no residuals to take to the third step—this is the extreme case of perfect multicollinearity.

Specifying the Regression Model

Suppose an expert wants to analyze the salaries of women and men at a large publishing house to discover whether a difference in salaries between employees with similar years of work experience provides evidence of discrimination.[142] To begin with the simplest case, the salary in dollars per year, Y, represents the dependent variable to be explained, and X_1 represents the number of years of experience of the employee explanatory variable. The regression model would be written

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{4}$$

In equation (4), β_0 and β_1 are the parameters to be estimated from the data, and ε is the error term. The parameter β_0 is the average salary of all employees with no experience. The parameter β_1 measures the average effect of an additional year of experience on the average salary of employees.

Fitted Regression Line

Once the parameters in a regression equation have been estimated, the fitted values for the dependent variable can be calculated. If the estimated regression parameters, or regression coefficients, for the model in equation (3) are denoted as $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, the fitted values for Y, denoted $\hat{Y}$, are

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \tag{5}$$

Figure 5 illustrates this for the example involving a single explanatory variable. The data are shown as a scatter of points; salary is on the vertical axis, and years of experience is on the horizontal axis. The estimated regression line is drawn through the data points. It is given by

$$\hat{Y} = \$15{,}000 + \$2{,}000 X_1 \tag{6}$$

Thus, the fitted value for the salary of individual i, who has X_{1i} years of experience, will be given by

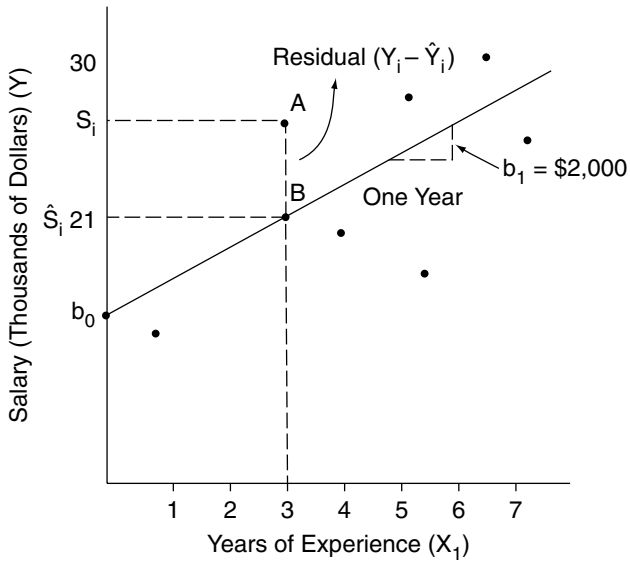
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} \text{ (Point B)} \tag{7}$$

142. The regression results used in this example are based on data for 1,715 men and women; these results were used by the defense in a sex-discrimination case against the *New York Times* that was settled in 1978. Professor Orley Ashenfelter, Department of Economics, Princeton University, provided the data.

The intercept of the regression model represents the average value of the dependent variable when the explanatory variable or variables are equal to 0; the estimated intercept b_0 is shown on the vertical axis in Figure 5. Similarly, the slope of the line measures the (average) change in the dependent variable associated with a unit increase in an explanatory variable; the estimated slope b_1 also is shown. In equation (6), the intercept \$15,000 indicates that employees with no experience earn \$15,000 per year. The slope parameter implies that each year of experience adds \$2,000 to an "average" employee's salary.

Now, suppose that the salary variable is related simply to the sex of the employee. The relevant indicator variable, often called a dummy variable, is X_2, which is equal to 1 if the employee is male, and 0 if the employee is female. Suppose the regression of salary Y on X_2 yields the following result: $Y = \$30{,}449 + \$10{,}979X_2$. The coefficient \$10,979 measures the difference between the average salary of men and the average salary of women.[143]

Figure 5. Goodness of fit.



143. To understand why, note that when $X_2 = 0$, the average salary for women is $\$30{,}449 + \$10{,}979 \times 0 = \$30{,}449$. Correspondingly, when $X_2 = 1$, the average salary for men is $\$30{,}449 + \$10{,}979 \times 1 = \$41{,}428$. The difference—\$41,428 – \$30,449—is \$10,979.

Regression residuals

Recall that for each data point, the regression residual is the difference between the actual value of the dependent variable and part of that value that is attributable to the explanatory variables. After a regression model has been estimated, the estimated regression residual ($\hat{\varepsilon}$) is the difference between the actual values and the fitted or predicted values from the estimated model. Suppose, for example, that we are studying an individual with three years of experience and a salary of $27,000. According to the estimated regression line in Figure 5, the predicted salary of an individual with three years of experience is $21,000. (This can also be interpreted as the expected salary that an individual with three years of experience would receive, since all the unexplained terms in the residual should on average take a value of 0.) Because the individual's salary is $6,000 higher than the predicted salary from the model, the estimated residual (the individual's salary minus the predicted salary, based on the estimated coefficients) is $6,000. In general, the estimated residual $\hat{\varepsilon}$ associated with a data point, such as Point A in Figure 5, is given by $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}$. Each data point in the figure has an estimated residual, which is the error made by the least-squares regression method for that individual.

Interaction terms

Models with interaction terms allow for the possibility that the effect of an explanatory variable on the dependent variable may vary in magnitude as the level of other explanatory variables changes. As a starting point, assume that the regression model takes the following form:

$$Y = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{EXP} + \varepsilon \tag{8}$$

where Y is annual salary, MALE is equal to 1 for men and 0 for women, EXP represents years of job experience, and ε is the error term. The coefficient β_1 measures the difference in average salary (across all experience levels) between men and women. In essence the inclusion of a dummy variable such as MALE allows for a parallel vertical shift in the regression line. The shift is parallel because the slope coefficient β_2, which measures the effect of experience on salary, is the same for both males and females.

Now suppose that the regression model is expanded to take on the following form:

$$Y = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{EXP} + \beta_3 \text{EXP} \times \text{MALE} + \varepsilon \tag{9}$$

In this model, the coefficient β_1 measures the difference in average salary (across all experience levels) between men and women for employees with no experience. The coefficient β_2 measures the effect of experience on salary for women (when MALE = 0), and the interaction coefficient β_3 measures the difference in the effect of experience on salary between men and women. It follows, for example, that the effect of one year of experience on salary for women is β_2, whereas the comparable effect for men is $\beta_2 + \beta_3$.[144]

Interpreting regression results

To explain how regression results are interpreted, we can expand the earlier example associated with Figure 5 to consider the possibility of an additional explanatory variable—the square of the number of years of experience, X_3. The X_3 variable is designed to capture the fact that for most individuals, salaries increase with experience, but at a faster rate for people who first enter the labor market (with low levels of experience), and a lower rate for people with more experience. Thus, we might expect the estimated coefficient of X_1 to be positive and the estimated coefficient of X_3 to be negative.[145]

The estimated regression line using the third additional explanatory variable, as well as the first explanatory variable for years of experience (X_1) and the dummy variable for male gender (X_2), is

$$\hat{Y} = \$14{,}085 + \$2{,}323X_1 + \$1{,}675X_2 - \$36X_3$$

The importance of including relevant explanatory variables in a regression model is illustrated by the change in the regression results after the X_3 and X_1 variables are added. The coefficient on the variable X_2 measures the difference in the salaries of men and women while controlling for the linear and quadratic effects of experience. The differential of $1,675 is substantially lower than the previously measured differential of $10,979. Clearly, failure to control for job experience in this example leads to an overstatement of the difference in salaries between men and women.

Now consider the interpretation of the explanatory variables for experience, X_1 and X_3. The positive sign on the X_1 coefficient shows that salary increases with experience. The negative sign on the X_3 coefficient indicates that the rate of

144. Estimating a regression in which there are interaction terms for all explanatory variables, as in equation (9), is essentially the same as estimating two separate regressions, one for men and one for women.

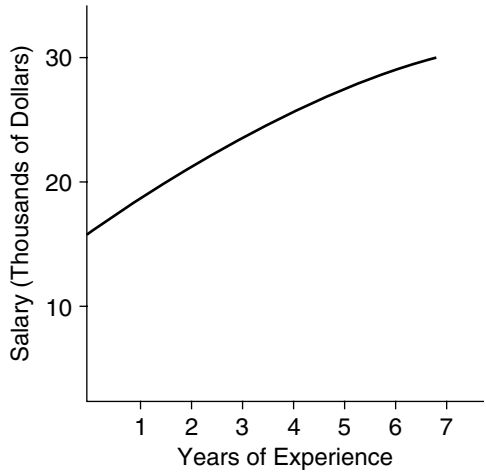
145. A model that includes a variable and its square is often referred to as a *quadratic* specification. Thus, the equation on this page would be referred to as a model that includes controls for "a quadratic in experience."

salary increase diminishes with experience (as expected). To determine the combined effect of the variables X_1 and X_3, some simple calculations can be made.[146] For example, consider how the average salary of women ($X_2 = 0$) changes with the level of experience. As experience increases from zero to one year, the average salary increases by \$2,287 (\$2,323 – \$36), from \$14,085 to \$16,372. However, women with two years of experience earn only \$2,215 more than women with one year of experience, and women with three years of experience earn only \$2,143 more than women with two years.[147] Figure 6 illustrates the results: The regression line shown is for women's salaries; the corresponding line for men's salaries would be parallel and \$1,675 higher.

The problem of omitted variables

A major problem confronting the interpretation of regression models is the possibility that some of the unobserved or omitted factors included in the error term are in fact correlated with the included variables. For example, consider an analysis of the effect of having attended a prestigious law school on the earnings of attorneys 10 years after completion of their law degrees. The available dataset

Figure 6. Regression slope for women's salaries.



146. More formally, consider a model that relates a dependent variable to some variable Z with a coefficient of β_1 and to its square (Z^2) with a coefficient of β_2 (i.e., $Y = \beta_1 Z + \beta_2 Z^2 + \ldots$). The marginal effect of a small increase in Z on the predicted value of the dependent variable is $\beta_1 + 2\beta_2 Z$, which is the derivative of the quadratic expression.

147. These numbers can be calculated by substituting different values of X_1 and X_3 in the previous equation.

includes a sample of students who attended different law schools, a measure (Y) of their annual salary in the tenth year post-graduation, an indicator (X_1) for having attended a "top 20" law school; an indicator for female gender (X_2), and an indicator (X_3) for whether their undergraduate degree was in a STEM (Science, Technology, Engineering, and Mathematics) field or not. The expert intends to interpret the coefficient on attending a top 20 school as the causal effect of the superior training and network opportunities afforded by such a school, as part of litigation over the admission rules used by certain schools.

The proposed regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{10}$$

Included in the error term ε are all the other unmeasured factors affecting salary, apart from the three covariates. Some of these factors, such as interest in pursuing a public interest law career, may negatively affect salary, while others, like ambition or ability, may positively affect salary. Moreover, it is likely that the average value of ε is different for students who attended a top 20 program versus those who did not. How does this affect the estimates of the coefficient β_1?

It turns out there is a very simple answer, at least conceptually. Consider the *hypothetical* regression model relating ε to the three covariates:

$$\varepsilon = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \xi \tag{11}$$

Here, the coefficients λ_1, λ_2, and λ_3 represent the effects of attending an elite school, being female, or having a STEM undergraduate degree on the combination of omitted factors that are included in ε. For example, if $\lambda_1 > 0$, then attending an elite school is associated, on average, with a more positive combination of omitted factors. Note that if the error term in equation (10) is truly random, and independent of the included covariates, then all the λ coefficients in (11) will be zero. Alternatively, if X_1 were randomly determined by an admissions lottery, then λ_1 would be zero, since there is no information in a randomized variable that can be used to predict other variables.

When the regression model (10) is estimated by OLS, the coefficients of the included covariates will incorporate both the true causal effects of the covariates (i.e., the β's), and the effect of the covariates in predicting ε (i.e., the λ's):

$$Y = (\beta_0 + \lambda_0) + (\beta_1 + \lambda_1) X_1 + (\beta_2 + \lambda_2) X_2 + (\beta_3 + \lambda_3) X_3 + \xi \tag{12}$$

Notice that the error in this model is the error from the hypothetical regression model (11): this is the part of ε that cannot be predicted by the included covariates. Equation (12) says that in estimating a multiple regression of salaries on elite school status, gender, and STEM major, the coefficient associated with attending an elite school is actually $\beta_1 + \lambda_1$, reflecting the fact that attending an

elite law school has a direct effect on salary, β_1, holding constant the other controls and all the unobservable factors in ε, plus an indirect effect arising because on average people who attend an elite law school have different values of ε. This second term, λ_1, is known as omitted-variables bias.

To summarize, if one estimates a multiple regression model relating an outcome to a set of observed covariates, the estimated coefficients will be estimates of the combined effect of the causal coefficients in the intended model *and* the coefficients that arise in trying to predict whatever is left in the error term in this model. This happens because the actual regression coefficients are algorithmically selected to form the best possible estimate of *Y*, given the observed covariates. In choosing the coefficients, the OLS algorithm *does not distinguish* between the intended causal effects in the model and the potential role of the covariates in predicting the residual component in the model.

Although equations (10)–(12) provide a useful framework for thinking about omitted-variables biases, it is important to keep in mind that the hypothetical regression (11) cannot be estimated, because we cannot actually see the value of ε. Nevertheless, analysts often present detailed discussions of the likely sign (and less commonly, the magnitude) of the coefficients in the hypothetical regression. If it can be argued, for example, that λ_1 is almost surely positive, then one can infer that the estimated effect of attending an elite law school from a model based on (10) is almost surely upward biased as an estimate of β_1. In addition, sometimes information from another sample can be used to help assess the signs and magnitudes of omitted-variables biases. For example, suppose information from a previous study shows that higher LSAT scores are associated with higher salaries and that information is available on the average LSAT scores of students attending different law schools. From these sources it might be possible to make an educated guess about a plausible magnitude for λ_1.

Causal Modeling

Pre/Post Comparisons

Concerns over omitted-variables bias in regression models have led to the development of alternative regression-based approaches for estimating the causal effect of a variable of interest. These approaches typically focus on situations where the value of the explanatory variable of interest changes discretely *for a known reason* (e.g., a political decision).[148] For example, suppose one is interested in understanding the effect of more intensive policing on crime. One might then focus on cases where a new policing program (e.g., New York City's "stop and frisk" program) is introduced. If data on crime rates are available from both *before* and *after*

148. Sometimes these changes are referred to as *quasi-experiments* or *natural experiments*.

the introduction, a pre/post estimate of the effect of the policy can be constructed from the change in the average number of crimes per day following the implementation of the new policy.

In a pre/post research design (and in related difference-in-differences designs), it is necessary to distinguish between the two separate dimensions over which the data are observed. One dimension is across different units, indexed by the subscript i (e.g., in a study of policing and crime, the units of observation may be police precincts); the other is over time (e.g., different months or years), indexed by the subscript t. Using this notation, let Y_{it} represent the outcome of interest observed for unit i in period t. Assume that a new program or intervention is initiated at time $t = 1$. In earlier periods ($t = 0$, $t = -1$, $t = -2$, . . .), the program was absent; in all periods on or after $t = 1$, the program is in place. A very simple model for Y_{it} is one that includes a constant and an indicator for post-intervention periods, $Post_t = 0$ if $t \leq 0$, and $Post_t = 1$ if $t \geq 1$:

$$Y_{it} = \beta_0 + \beta_1 Post_t + \varepsilon_{it}. \tag{13}$$

This simple model implies that in all periods up to period 0,

$$Y_{it} = \beta_0 + \varepsilon_{it} \tag{14}$$

whereas in all later periods $t = 1, 2, \ldots,$

$$Y_{it} = \beta_0 + \beta_1 + \varepsilon_{it}. \tag{15}$$

Setting $t = 1$ in equation (15) and setting $t = 0$ in equation (14) and subtracting, the change in the outcome for unit i from period 0 to period 1 is

$$\Delta Y_i = Y_{i1} - Y_{i0} = \beta_1 + \varepsilon_{i1} - \varepsilon_{i0}. \tag{16}$$

Treating $\varepsilon_{i1} - \varepsilon_{i0}$ as a residual, equation (16) is the simplest possible regression model, with a constant term and no other covariates. The OLS estimate of β_1 in such a model is exactly the mean change in the outcome across all the units in the sample, that is,

$$\hat{\beta}_1 = Mean(\Delta Y_i) \tag{17}$$

where $Mean(Y)$ denotes the mean value of the variable Y in the sample.

To interpret equation (17), note that $Mean(\Delta Y_i) = Mean(Y_{i1}) - Mean(Y_{i0})$. In a pre/post design, the mean outcome in the post period is compared to the mean outcome in the pre period. If the mean value of the outcome does not change after the intervention, the estimated effect of the intervention is 0. If the mean outcome rises (falls), then $\hat{\beta}_1$ will be positive (negative). Conceptually, a pre/post

design is using the mean of the outcome in the pre-intervention period as an estimate of what the mean *would have been* in the absence of the intervention. This is known as the counterfactual mean in applied economics, or the but-for mean in legal discussions. The difference between the actual mean in the post-intervention period and the counterfactual mean is the estimated effect of the intervention.

If data were only available from the period after the start of the intervention, an alternative approach would be to try to find another set of units that have not adopted the intervention and use the mean of the outcomes for this "comparison group" as a counterfactual. In the policing example, suppose that some precincts adopt the new program, and some do not. Then one might consider using the average crime rates for precincts that did not adopt the policy as the counterfactual for those that did.[149] A concern with this alternative approach is that there may be differences between units that adopted the new program and units that did not, raising the possibility of omitted-variables bias. In a pre/post design, the counterfactual is based on the average outcomes *of the same units* before they implemented the intervention. In some cases, this counterfactual will be more compelling.

The potential advantages of a pre/post design are illustrated by decomposing ε_{it}, the error term in equation (13), into two parts: a part α_i that is constant over time, representing the mean value of all the ε_{it}'s for unit i, and the deviation of ε_{it} from its mean value in period t,

$$\varepsilon_{it} = \alpha_i + \nu_{it} \tag{18}$$

For example, in a study of crime and policing where i indexes different precincts and t has two values, $t = 0$ and $t = 1$, the error ε_{it} from the model in equation (11) has the interpretation of the deviation in the crime rate in a precinct i in period t from the average crime rate in the city in the same period. One might expect certain precincts to have higher average crime rates in both periods, and others to have lower crime rates. Such differences in average rates would be reflected in the α_i component of equation (18), whereas unexplained fluctuations in crime from period to period in the same precinct would be reflected in the ν_{it} component. From equation (18) it follows that the difference in ε_{it} between period 0 and period 1, which is the error term in equation (16), is

$$\varepsilon_{i1} - \varepsilon_{i0} = \nu_{i1} - \nu_{i0} \tag{19}$$

Differencing eliminates the time-invariant component of ε_{it}, leaving a new error term that is less likely to be affected by omitted-variable biases, since all the

149. This would replace equation (13) with an alternative model: $Y_i = \beta_0 + \beta_1 Adopt_i + \varepsilon_i$, where $Adopt_i$ is an indicator variable equal to 1 for precincts that adopted the policy and 0 for those that did not. In such a model, the OLS estimate of the coefficient β_1 is $\hat{\beta}_1 = Mean(Y \mid Adopt_i = 1) - Mean(Y \mid Adopt_i = 0)$, where $Mean(Y|C)$ denotes the mean in the sample among units for which condition C is true.

permanent factors that differ across units are removed through the differencing process.

Nevertheless, a simple pre/post design has an important limitation: *any change* that affects the average value of the outcome between period 0 and period 1 is attributed to the intervention. In many situations, the mean value of an outcome changes over time for reasons that have nothing to do with the intervention or change in policy under study. One way to check whether such factors are present is to construct average differences in the outcome for earlier periods and verify that these are all very close to zero. For example, equation (13) implies that the *lagged difference* in outcomes $\Delta Y_i^{(-1)}$ is determined by

$$\Delta Y_i^{(-1)} = Y_{i0} - Y_{i(-1)} = \varepsilon_{i0} - \varepsilon_{i(-1)} \ . \tag{20}$$

The same arguments that support the contention that the error term $\varepsilon_{i1} - \varepsilon_{i0}$ in equation (16) has a mean value of 0, so that $Mean(\Delta Y_i)$ provides a good estimate of β_1, imply that the error term in equation (20) also has a mean value of 0. Thus, $Mean(\Delta Y_i^{(-1)})$ should be very close to 0. Indeed, the mean values of the changes in the outcome for all periods where there was no change in policy should be close to 0. In the absence of evidence that this is true, a simple pre/post design is not very credible.

Difference in Differences

A natural way to address the presence of unmeasured factors that are changing over time and confounding the interpretation of a simple pre/post design is to compare the change in the outcome for units where the new program was adopted to the change in the outcome over the same period for units that did not adopt the program. This is the difference-in-differences approach.

To spell this out, call the treatment group the set of units where the program is adopted between period 0 and period 1; call the other set of units the comparison group, and let T_i be an indicator variable that is equal to 1 for units in the treatment group. Suppose that data are only observed in periods 0 and 1. Then, the difference-in-differences model can be written as a multiple linear regression model of the form

$$Y_{it} = \beta_0 + \beta_1 T_i \times Post_t + \beta_2 T_i + \beta_3 Post_t + \varepsilon_{it} \tag{21}$$

The key coefficient in this model is β_1, the coefficient on the interaction between the indicator for the treatment group and the indicator for period 1 (the post period).

To understand this equation, notice that for units with $T_i = 0$, it implies

$$Y_{it} = \beta_0 + \beta_3 Post_t + \varepsilon_{it} \tag{22}$$

Thus, the change in outcomes for units in the comparison group is

$$\Delta Y_i = Y_{i1} - Y_{i0} = \beta_3 + \varepsilon_{i1} - \varepsilon_{i0} \tag{23}$$

The inclusion of the term $\beta_3 Post_t$ in equation (22) captures the possibility that even in the comparison group, unobserved factors can change between period 0 and period 1. Assuming the mean value of $\varepsilon_{i1} - \varepsilon_{i0}$ is 0 for comparison units, the expected average change in the outcome in the comparison group will be β_3.[150] For units in the treatment group, equation (21) implies

$$Y_{it} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Post_t + \varepsilon_{it} \tag{24}$$

Thus, the change in outcomes for units in the treatment group is

$$\Delta Y_i = (\beta_1 + \beta_3) + \varepsilon_{i1} - \varepsilon_{i0} \tag{25}$$

Assuming that the mean value of $\varepsilon_{i1} - \varepsilon_{i0}$ is 0 for units in the treatment group, the expected average change in the outcome in the treatment group will be $\beta_1 + \beta_3$, which is a combination of the change experienced by the comparison group, β_3, and the effect of the program, β_1. In fact, the OLS estimate of β_1 in the difference-in-differences model (21) is exactly

$$\hat{\beta}_1 = Mean(\Delta Y_i \mid T_i = 1) - Mean(\Delta Y_i \mid T_i = 0) \tag{26}$$

where $Mean(\Delta Y_i \mid T_i = 1)$ denotes the mean in the subsample with $T_i = 1$ (i.e., among the treatment group), and $Mean(\Delta Y_i \mid T_i = 0)$ denotes the mean in the subsample with $T_i = 0$ (i.e., among the comparison group).

The critical assumption in a difference-in-differences design is that in the absence of the program, the mean change in outcomes for the treatment group and comparison groups would have been the same. Under that assumption, the mean change for the comparison group forms a counterfactual for the change that would have occurred in the treatment group in the absence of the program. As noted previously, this assumption is sometimes referred to as a parallel trends assumption. With only two periods of data, it cannot be directly evaluated. If there are additional periods of data, however, parallel trends can be assessed by forming differences in differences for other periods that do not overlap with the start of the program. These should all be close to zero. Visually, a graph of the mean outcomes for the treatment group over a sequence of periods (for example, $t = -2, -1, 0, 1, 2$) should show that the the gap between the treatment means and

150. To be precise, the "expected average change" means the mathematical expectation of the sample average change in the outcome.

comparison means is stable prior to the program date, then shifts discretely between period 0 and 1, reflecting the treatment effect, then stabilizes again.

Another way of thinking about the validity of a difference-in-differences specification is to ask whether the choice of the pre-treatment period matters. For example, consider an evaluation of a state law that liberalizes state divorce law. A difference-in-differences approach to measuring the effect of the law would be to compare the change in divorce rates from some year prior to the law to the year after the law changed in the "treatment" state (the state that passed the new law) relative to the change in divorce rates in a "comparison" state. If the assumptions of a difference-in-differences specification are correct, then the relative change in divorce rates in the treated state relative to the comparison state should not depend on the particular year that is selected as the benchmark prior to the law change. Under those assumptions, the difference in divorce rates in the treated and comparison states can be expected to be similar in all the years prior to the new law. As a result, choosing one particular year as a benchmark versus another will not matter. Graphically, this requires that the year-to-year changes in divorce rates in the treated state and the comparison state are the same, and a plot of the divorce rates prior to the law change would appear as two parallel lines. This is the origin of the term *parallel trends*.

A common scenario where the parallel trends assumption may fail is one in which the timing of the program is correlated with the trend in recent outcomes for the treatment group. For example, Ashenfelter (1978) noted that people who entered government retraining programs had *temporarily depressed* earnings in the period before entry, driven by recent job losses or other problems—the Ashenfelter dip.[151] Building on equation (18), this suggests that program participants may be particularly likely to have large negative values of ν_{i0} (the period-specific part of their earnings residual in the period just prior to the program). In that case, the mean value of the error term $\varepsilon_{i1} - \varepsilon_{i0} = \nu_{i1} - \nu_{i0}$ in equation (25) will be positive, implying that

$$Mean(\Delta Y_i | T_i = 1) = \beta_1 + \beta_3 + Mean(\varepsilon_{i1} - \varepsilon_{i0} | T_i = 1) > \beta_1 + \beta_3 \qquad (27)$$

causing a difference-in-differences estimator to overstate the effect of the program. The possibility of such pre-trends can be evaluated empirically by examining the mean outcomes of the treatment and comparison groups in the periods before the start of the program and making sure that the parallel trends assumption is valid.

151. Ashenfelter, *supra* note 94.

Instrumental Variables

An alternative method that is widely used to address omitted variables problems in regression models is instrumental variables (IV). Consider the case where an expert has posited a model relating an outcome Y to an observed explanatory variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{28}$$

For clarity, subscripts indicate the different units (indexed by i) in the sample. The expert wants to give a causal interpretation of the coefficient β_1. Specifically, in the expert's model, β_1 represents the effect of a unit change in X on the average or expected value of the outcome, holding constant all other unmeasured/unobserved determinants of Y. Such an equation is known as the structural model in IV and related settings. The problem is that, if this structural equation is estimated by OLS, the estimate $\hat{\beta}_1$ will incorporate both the desired causal effect and any potential omitted variable bias that arises because X_i is potentially correlated with some of the factors included in the error ε_i. This omitted-variable bias is sometimes called a simultaneity bias or endogeneity bias—particularly in cases where there is feedback from Y_i to X_i, as discussed in the section titled "Instrumental Variables Design" above. Whatever the source of the potential correlation between X_i and ε_i, however, it can be analyzed in the omitted variables framework of equations (10)–(12).

The objective of IV estimation is to isolate the part of the variation in X_i attributable to variation in some other variable, Z_i (the instrumental variable or the instrument), and use only that variation in estimating the effect of X_i on Y_i. An appropriate instrumental variable has to satisfy three assumptions: (1) it has an effect on X_i, so there is a component of X_i that can be attributed to Z_i; (2) it is uncorrelated with ε_i, the unobserved determinants of Y_i; and (3) it has no direct effect on the outcome Y_i. An example of an instrumental variable that satisfies these requirements is a randomly assigned lottery outcome. For example, in a study of the effect of attending a charter school on test scores, the analysis showed that if access to over subscribed charter schools is determined by lottery, then an appropriate instrumental variable is an indicator for whether a student was assigned to the charter school by the admission lottery.[152]

The IV method involves estimating two regression models. The first-stage model is a linear regression relating the variable of interest (X_i) to the instrumental variable:

$$X_i = \pi_0 + \pi_1 Z_i + u_i \tag{29}$$

152. *See generally* Chabrier et al., *supra* note 104.

From this model it is possible to form the predicted values of X_i based only on Z_i:

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i \tag{30}$$

The second-stage model is a linear regression of the outcome on the predicted values of X_i:

$$Y_i = \beta_0^{iv} + \beta_1^{iv} \hat{X}_i + \eta_i \tag{31}$$

The OLS estimate of the coefficient of $\hat{X}_i$ in the second-stage model, $\hat{\beta}_1^{iv}$, is the IV estimate of β in the structural model (28). Note that since $\hat{X}_i$ only depends on Z_i, the second-stage equation measures the effect of the part of $\hat{X}_i$ that is determined by Z_i on the outcome variable.

The IV estimate can be derived in another way. Substituting the first-stage equation (29) into the structural equation (28) yields an equation relating Y to Z:

$$Y_i = \underbrace{(\beta_0 + \beta_1 \pi_0)}_{\delta_0} + \underbrace{\beta_1 \pi_1}_{\delta_1} Z_i + \underbrace{(\varepsilon_i + \beta_1 u_i)}_{\xi_i} = \delta_0 + \delta_1 Z_i + \xi_i \tag{32}$$

This equation, known as the reduced-form model, can be estimated by OLS without concern about omitted variables bias because, under the assumption that Z_i is a valid instrumental variable, it is uncorrelated with ξ_i.[153] But equation (32) shows that the coefficient of the instrumental variable in the reduced-form model is $\delta_1 = \beta_1 \pi_1$. This occurs because under assumption #3 for a valid instrumental variable, the *only way* that Z_i affects Y_i is indirectly, through its effect on X_i. Since Z_i affects X_i with a coefficient of π_1 in the first-stage model, and X_i affects Y_i with a coefficient of β_1 in the structural model, the reduced-form effect of Z_i on Y_i is the product of these two coefficients. This reasoning suggests that one could obtain an estimate of β_1 by dividing the estimated reduced-form coefficient, $\hat{\delta}_1$, by the estimated first-stage coefficient $\hat{\pi}_1$. And in fact this ratio is exactly the IV estimate:

$$\hat{\beta}_1^{iv} = \frac{\hat{\delta}_1}{\hat{\pi}_1} \tag{33}$$

The expression for the IV estimate in equation (33) is particularly insightful in the case where X_i is a dummy variable indicating participation in a program or

153. Under assumption #2 for a valid instrumental variable, Z_i is uncorrelated with ε_i. Moreover, the error in the first-stage equation u_i, is necessarily uncorrelated with the explanatory variable in that model, by a fundamental property of least squares. Thus Z_i is uncorrelated with the composite error in the reduced-form model, $\xi_i = \varepsilon_i + \beta_1 u_i$.

intervention (for example, enrollment in a charter school), and Z_i is also a dummy variable, indicating a condition that partly determines program participation (for example, whether student i was assigned to the charter school by lottery). In this case, the first-stage coefficient $\hat{\pi}_1$ is the difference in program participation rates between units with Z_i equal to 0 or 1:

$$\hat{\pi}_1 = Mean(X_i | Z_i = 1) - Mean(X_i | Z_i = 0), \tag{34}$$

and the reduced-form coefficient $\hat{\delta}_1$ is the difference in the mean of the outcome variable (e.g., an end-of-year test score) for the same two groups:

$$\hat{\delta}_1 = Mean(Y_i | Z_i = 1) - Mean(Y_i | Z_i = 0). \tag{35}$$

Thus, equation (33) implies that the IV estimator for the effect of program participation on the outcome is

$$\hat{\beta}_1^{iv} = \frac{Mean(Y_i | Z_i = 1) - Mean(Y_i | Z_i = 0)}{Mean(X_i | Z_i = 1) - Mean(X_i | Z_i = 0)} \tag{36}$$

The IV estimator takes the difference in mean outcomes between the group with $Z_i = 1$ and the group with $Z_i = 0$, and divides this by the difference in the fraction who participated in the program when Z_i was 1 or 0. Intuitively, the IV estimator is simply creating a *per unit* estimate of the treatment effect of the intervention, based on the differences between groups with different values of Z_i.

Determining the Precision of the Regression Results

Least squares regression algorithms provide not only parameter estimates that indicate the direction and magnitude of the effect of a change in the explanatory variable on the dependent variable, but also an estimate of the reliability of the parameter estimates and a measure of the overall goodness of fit of the regression model. Each of these factors is considered in turn.

Standard Errors of the Coefficients and t-Statistics

Estimates of the true but unknown parameters of a regression model are numbers that depend on the sample of observations under study. If a different sample

were used, a different estimate would be calculated.[154] If the expert continued to collect more and more samples and generated additional estimates, as might happen when new data became available over time, the estimates of each parameter would follow a probability distribution. This probability distribution can be summarized by a mean and a measure of dispersion around the mean, a standard deviation, which usually is referred to as the standard error of the coefficient, or the standard error (SE).[155]

Suppose, for example, that an expert is interested in estimating the average price paid for a gallon of unleaded gasoline by consumers in a particular geographic area of the United States at a particular point in time. The mean price for a sample of 10 gas stations might be $1.25, while the mean for another sample might be $1.29, and the mean for a third, $1.21. On this basis, the expert also could calculate the overall mean price of gasoline to be $1.25 and an estimate of the standard deviation to be $0.04.

Least-squares regression generalizes this result, by calculating means whose values depend on one or more explanatory variables. The standard error of a regression coefficient tells the expert how much parameter estimates are likely to vary from sample to sample. The greater the variation in parameter estimates from sample to sample, the larger the standard error and consequently the less reliable the regression results. Small standard errors imply results that are likely to be similar from sample to sample, whereas results with large standard errors show more variability.

Under appropriate assumptions, the ordinary least-squares estimators provide "best" determinations of the true underlying parameters.[156] In fact, OLS has several desirable properties. First, assuming that the error term in the regression model is uncorrelated with the covariates, least squares estimators are unbiased. Intuitively, this means that if the regression were calculated repeatedly with different samples, the average of the many estimates obtained for each coefficient would be the true parameter. Second, under the same assumption, least-squares estimators are consistent; if the sample were very large, the estimates obtained would come close to the true parameters. Third, least squares is efficient, in that its estimators have the smallest variance among all (linear) unbiased estimators.

If the further assumption is made that the probability distribution of each of the error terms is known, statistical statements can be made about the precision

154. The least squares formula that generates the estimates is called the least squares estimator, and its values vary from sample to sample.

155. *See* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Randomized Controlled Experiments" section, in this manual.

156. The necessary assumptions of the regression model include (a) the model is specified correctly, (b) errors associated with each observation are drawn randomly from the same probability distribution and are independent of each other, (c) errors associated with each observation are independent of the corresponding observations for each of the explanatory variables in the model, and (d) no explanatory variable is correlated perfectly with a combination of other variables.

of the coefficient estimates. For relatively large samples (often, thirty or more data points will be sufficient for regressions with a small number of explanatory variables), the probability that the estimate of a parameter lies within an interval of 1.96 standard errors around the true parameter is approximately .95, or 95%. A frequent, although not always appropriate, assumption in statistical work is that the error term follows a normal distribution, from which it follows that the estimated parameters are normally distributed. The normal distribution has the property that the area within 1.96 standard errors of the mean is equal to 95% of the total area. Note that the normality assumption is not necessary for least squares to be used, because most of the properties of least squares apply regardless of normality.

In general, for any parameter estimate $\hat{\beta}$, the expert can construct an interval around $\hat{\beta}$ such that if the sampling procedure were conducted 100 times, approximately 95 of the resulting intervals would be expected to contain the true population mean. The 95% confidence interval[157] is given by[158]

$$\hat{\beta} \pm 1.96(\text{SE of } \hat{\beta}) \tag{37}$$

The expert can test the hypothesis that a parameter is equal to 0 (often stated as testing the null hypothesis) by looking at its *t*-statistic, which is defined as

$$t = \hat{\beta}/\text{SE}(\hat{\beta}) \tag{38}$$

If the *t*-statistic is less than 1.96 in magnitude, the 95% confidence interval around $\hat{\beta}$ must include 0.[159] Because this means that the expert cannot reject the hypothesis that β equals 0, the estimate, whatever it may be, is said to be not statistically significant. Conversely, if the *t*-statistic is greater than 1.96 in absolute value, the expert concludes that the true value of β is unlikely to be 0 (intuitively, $\hat{\beta}$ is "too far" from 0 to be consistent with the true value of β being 0). In this case, the expert rejects the hypothesis that β equals 0 and calls the estimate statistically significant. If the null hypothesis β equals 0 is true, using a 95% confidence level will cause the expert to falsely reject the null hypothesis 5% of the time. Consequently, results often are said to be significant at the 5% level.[160]

157. Confidence intervals are used commonly in statistical analyses because the expert can never be certain that a parameter estimate is equal to the true population parameter.

158. If the number of data points in the sample is small (perhaps less than 30), the 1.96 number must be adjusted upward.

159. The *t*-statistic applies to any sample size. As the sample size increases, the underlying distribution, which is the source of the *t*-statistic (Student's *t*-distribution), approximates the normal distribution.

160. A *t*-statistic of 2.57 in magnitude or greater is associated with a 99% confidence level, or a 1% level of significance, that includes a band of 2.57 standard deviations on either side of the estimated coefficient.

As an example, consider a more complete set of regression results associated with the salary regression described earlier:

$$\begin{array}{rcccc} \hat{Y} = & \$14{,}085 + & \$2{,}323X_1 + & \$1{,}675X_2 - & \$36X_3 \\ & (1{,}577) & (140) & (1{,}435) & (3.4) \\ t = & 8.9 & 16.5 & 1.2 & -10.8 \end{array} \quad (39)$$

The standard error of each estimated parameter is given in parentheses directly below the estimated coeffcient, and the corresponding *t*-statistics appear below the standard error values.

Consider the coefficient on the dummy variable X_2, representing the sex of a worker. It indicates that $1,675 is the best estimate of the mean salary difference between men and women. However, the standard error of $1,435 is large in relation to its coefficient $1,675. Because the standard error is relatively large, the range of possible values for measuring the true salary difference, the true parameter, is great. In fact, a 95% confidence interval is given by the range negative $1,138 to positive $4,488. In other words, the expert can have 95% confidence that the true value of the coefficient lies between –$1,138 and $4,488. Because this range includes 0, the effect of sex on salary is said to be insignificantly different from 0 at the 5% level. The *t* value of 1.2 is equal to $1,675 divided by $1,435. Because this *t*-statistic is less than 1.96 in magnitude (a condition equivalent to the inclusion of a 0 in the above confidence interval), the sex variable again is said to be an insignificant determinant of salary at the 5% level of significance.

Note also that experience is a highly significant determinant of salary, because both the X_1 and the X_3 variables have *t*-statistics substantially greater than 1.96 in magnitude. More experience has a significant positive effect on salary, but the size of this effect diminishes significantly with experience.

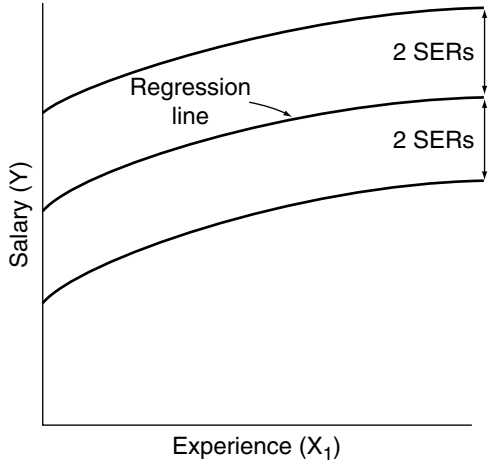
Goodness of Fit

Reported regression results usually contain not only the point estimates of the parameters and their standard errors or *t*-statistics, but also other information that tells how closely the regression line fits the data. One statistic, the standard error of the regression (SER), is an estimate of the overall size of the regression residuals.[161] An SER of 0 would occur only when all data points lie exactly on the regression line—an extremely unlikely possibility. Other things being equal, the larger the SER, the poorer the fit of the data to the model.

161. More specifically, it is a measure of the standard deviation of the regression error ε. It sometimes is called the root mean square error of the regression line.

For a normally distributed error term, the expert would expect approximately 95% of the data points to lie within two SERs of the estimated regression line, as shown in Figure 7.

Figure 7. Standard error of the regression (SER).



R-squared (R^2) is a statistic that measures the percentage of variation in the dependent variable that is accounted for by all the explanatory variables.[162] Thus, R^2 provides a measure of the overall goodness of fit of the multiple regression equation. Its value ranges from 0 to 1. An R^2 of 0 means that the explanatory variables explain none of the variation of the dependent variable; an R^2 of 1 means that the explanatory variables explain all the variation. The R^2 is .56. This implies that the three explanatory variables explain 56% of the variation in salaries.

What level of R^2, if any, should lead to a conclusion that the model is satisfactory? Unfortunately, there is no clear-cut answer to this question because the magnitude of R^2 depends on the characteristics of the data being studied and whether the data vary over time or over individuals. Typically, an R^2 is low in cross-sectional studies in which differences in individual behavior are explained. It is likely that these individual differences are caused by many factors that cannot be measured. As a result, the expert cannot hope to explain most of the variation. In time-series studies, in contrast, the expert is explaining the movement of aggregates over time. Because most aggregate time series have substantial

162. The variation is the square of the difference between each *Y* value and the average *Y* value, summed over all the *Y* values.

growth, or trend, in common, it will not be difficult to "explain" one time series using another time series, simply because both are moving together. It follows as a corollary that a high R^2 does not by itself mean that the variables included in the model are the appropriate ones.

Courts should be reluctant to rely solely on a statistic such as R^2 to choose one model over another. Alternative procedures and tests are available.[163] When utilizing probit or logit models, the R^2 is especially problematic, since (with the dependent variable taking on only two values) predictions will be spread across the 0–1 interval with relatively few values typically being close to either 0 or 1. As a result, R^2 values will be relatively low whether the estimated model coefficients are informative or not. One more instructive measure is the percentage of correct classifications with the prediction being 1 if the predicted value is greater than .5 and 0 otherwise. An alternative measure is a quasi-*R*-squared measure based on the log likelihood of the estimated model, relative to that of a model with no covariates.

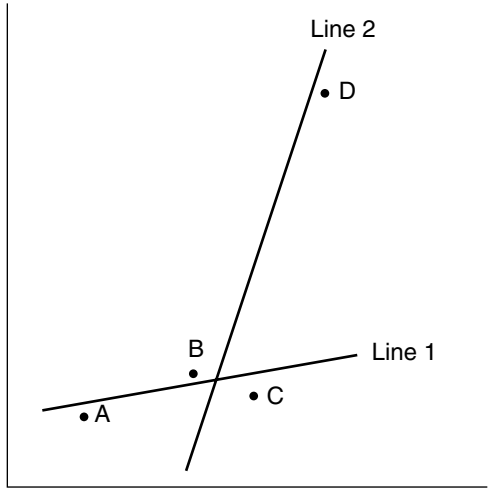
Sensitivity of Least-Squares Regression Results

The least-squares regression line can be sensitive to extreme data points. This sensitivity can be seen most easily in Figure 8. Assume initially that there are only three data points, A, B, and C, relating information about X_1 to the variable *Y*. The least-squares line describing the best-fitting relationship between Points A, B, and C is represented by Line 1. Point D is called an outlier because it lies far from the regression line that fits the remaining points. When a new, best-fitting least-squares line is re estimated to include Point D, Line 2 is obtained. Figure 8 shows that the outlier, Point D, is an influential data point, because it has a dominant effect on the slope and intercept of the least-squares line. Because least squares minimizes the sum of squared deviations, the sensitivity of the line to individual points sometimes can be substantial.[164]

163. These include *F*-tests and specification error tests. *See* Pindyck & Rubinfeld, *supra* note 23, at 88–95, 128–36, 194–98.

164. This sensitivity is not always undesirable. In some instances, it may be much more important to predict Point D when a big change occurs than to measure the effects of small changes accurately.

Figure 8. Least-squares regression.



What makes the influential data problem even more difficult is that the effect of an outlier may not be seen readily if deviations are measured from the final regression line. The reason is that the influence of Point D on Line 2 is so substantial that its deviation from the regression line is not necessarily larger than the deviation of any of the remaining points from the regression line.[165] Although they are not as popular as least squares, alternative estimation techniques that are less sensitive to outliers, such as robust estimation, are available. Alternatively, if the sample is relatively large, an expert may choose to remove the outliers before estimating the model.

165. The importance of an outlier also depends on its location in the dataset. Outliers associated with relatively extreme values of explanatory variables are likely to be especially influential. *See, e.g.*, Fisher v. Vassar College, 70 F.3d 1420, 1436 (2d Cir. 1995) (court required to include assessment of "service in academic community," because concept was too amorphous and not a significant factor in tenure review), *rev'd on other grounds*, 114 F.3d 1332 (2d Cir. 1997) (en banc). Conclusory statements of an outlier's impact are insufficient. *See In re* Nektar Therapeutics Sec. Litig., 34 F.4th 828, 836 (9th Cir. 2022) ("[t]he complaint does not allege with specificity what the Phase 1 EXCEL results would have been without outlier data").

A Hypothetical Example

Jane Thompson filed suit in federal court alleging that officials in the police department discriminated against her and a class of other female police officers in violation of Title VII of the Civil Rights Act of 1964, as amended. On behalf of the class, Officer Thompson alleged that she was paid less than male police officers with equivalent skills and experience. Both the plaintiff and the defendant used expert economists with econometric expertise to present statistical evidence to the court in support of their positions.

The plaintiff's expert pointed out that the mean salary of the 40 female officers was $30,604, whereas the mean salary of the 60 male officers was $43,077. To show that this difference was statistically significant, the expert put forward a regression of salary (SALARY) on a constant term and a dummy indicator variable (FEM) equal to one for each female and zero for each male. The results were as follows:

$$
\begin{array}{rll}
\text{SALARY} = & \$43{,}077 - & \$12{,}373 \times \text{FEM} \\
\text{St. Error} & (\$1{,}528) & (\$2{,}416) \\
p\text{-value} & <.01 & <.01 \\
R^2 = .22 & &
\end{array}
$$

The –$12,373 coefficient on the FEM variable measures the mean difference between male and female salaries. Because the standard error is approximately one-fifth of the value of the coefficient, this difference is statistically significant at the 5% (and indeed at the 1%) level. If this is an appropriate regression model (in terms of its implicit characterization of salary determination), one can conclude that it is highly unlikely that the difference in salaries between men and women is due to chance.

The defendant's expert testified that the regression model put forward was the wrong model because it failed to account for the fact that males (on average) had substantially more experience than females. The relatively low R^2 was an indication that there was substantial unexplained variation in the salaries of male and female officers. An examination of data relating to years spent on the job showed that the average male experience was 8.2 years, whereas the average for females was only 3.5 years. The defense expert then presented a regression analysis that added an additional explanatory variable (i.e., a covariate), the years of experience of each police officer (EXP).

The new regression results were as follows:

SALARY =	\$28,049 –	\$3,860 × FEM +	\$1,833 × EXP
St. Error	(\$2,513)	(\$2,347)	(\$265)
p-value	<.01	<.11	<.01
R^2 = .47			

Experience is itself a statistically significant explanatory variable, with a *p*-value of less than .01. Moreover, the difference between male and female salaries, holding experience constant, is only \$3,860, and this difference is not statistically significant at the 5% level. The defense expert was able to testify on this basis that the court could not rule out alternative explanations for the difference in salaries other than the plaintiff's claim of discrimination.

The debate did not end here. On rebuttal, the plaintiff's expert made three distinct points. First, whether \$3,860 was statistically significant or not, it was practically significant, representing a salary difference of more than 10% of the mean female officers' salaries. Second, although the result was not statistically significant at the 5% level, it was significant at the 11% level.

Third, and most importantly, the expert testified that the regression model was not correctly specified. Further analysis by the expert showed that the value of an additional year of experience was \$2,333 for males on average, but only \$1,521 for females. Based on supporting testimonial experience, the expert testified that one could not rule out the possibility that the mechanism by which the police department discriminated against females was by rewarding males more for their experience than females. The expert made this point clear by running an additional regression in which a further covariate was added to the model. The new variable was an interaction variable, measured as the product of the FEM and EXP variables. The regression results were as follows:

SALARY =	\$35,122 –	\$5,250 × FEM +	\$2,333 × EXP –	\$812 × FEM × EXP
St. Error	(\$2,825)	(\$3,347)	(\$265)	(\$185)
p-value	<.01	<.11	<.01	<.01
R^2 = .65				

The plaintiff's expert noted that for all males in the sample, FEM = 0, in which case the regression results are given by the equation

$$\text{SALARY} = \$35{,}122 + \$2{,}333 \times \text{EXP}$$

However, for females, FEM = 1, in which the corresponding equation is

$$\text{SALARY} = \$29{,}872 + \$1{,}521 \times \text{EXP}$$

It appears, therefore, that females are discriminated against not only when hired (i.e., when EXP = 0), but also in the reward they get as they accumulate more experience.

The debate between the experts continued, focusing less on the statistical interpretation of any one regression model, but more on the model choice itself, and not simply on statistical significance, but also to practical significance.

Glossary of Terms

alternative hypothesis. See hypothesis test.

association. The degree of statistical dependence between two or more events or variables. Events are said to be associated when they occur more frequently together than one would expect by chance.

bias. Any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values (i.e., the results are either too high or too low). A biased estimator of a parameter differs on average from the true parameter.

coefficient. An estimated regression parameter.

consistent estimator. An estimator that tends to become more and more accurate as the sample size grows.

correlation. A statistical means of measuring the linear association between variables. Two variables are correlated positively if, on average, they move in the same direction; two variables are correlated negatively if, on average, they move in opposite directions.

covariate. A variable that is possibly predictive of an outcome under study; an explanatory variable.

cross-sectional analysis. A type of analysis in which each data point is associated with a different unit of observation (e.g., an individual or a firm) measured at a particular point in time.

degrees of freedom (DF). The number of observations in a sample minus the number of estimated parameters in a regression model. A useful statistic in hypothesis testing.

dependent variable. The variable to be explained or predicted in a multiple regression model.

dummy variable. A variable that takes on only two values, usually 0 and 1, with one value indicating the presence of a characteristic, attribute, or effect (1), and the other value indicating its absence (0).

endogenous variable. A covariate for which there are unobserved factors that affect both the covariate and the dependent variable.

error term. A variable in a multiple regression model that represents the cumulative effect of several sources of modeling error.

estimate. The calculated value of a parameter based on the use of a particular sample.

estimator. The sample statistic that estimates the value of a population parameter (e.g., a regression parameter); its values vary from sample to sample.

ex ante forecast. A prediction about the values of the dependent variable that go beyond the sample; consequently, the forecast must be based on predictions for the values of the explanatory variables in the regression model.

explanatory variable. A variable that is associated with changes in a dependent variable.

ex post forecast. A prediction about the values of the dependent variable made during a period in which all values of the explanatory and dependent variables are known. Ex post forecasts provide a useful means of evaluating the fit of a regression model.

***F*-test.** A statistical test (based on an *F*-ratio) of the null hypothesis that a group of explanatory variables are jointly equal to 0. When applied to all the explanatory variables in a multiple regression model, the *F*-test becomes a test of the null hypothesis that R^2 equals 0.

feedback. When changes in an explanatory variable affect the values of the dependent variable, and changes in the dependent variable also affect the explanatory variable. When both effects occur at the same time, the two variables are described as being determined simultaneously.

fitted value. The estimated value for the dependent variable.

fixed-effects model. A regression model that includes a set of dummy variables that account for certain characteristics of individuals in the sample.

heteroscedasticity. When the error associated with a multiple regression model has a nonconstant variance; that is, the error values associated with some observations are typically high, while the values associated with other observations are typically low.

homoscedasticity. When the error associated with a multiple regression model has a constant variance.

hypothesis test. A statement about the parameters in a multiple regression model. The null hypothesis may assert that certain parameters have specified values or ranges; the alternative hypothesis would specify other values or ranges.

independence. When two variables are not correlated with each other (in the population).

independent variable. An explanatory variable that affects the dependent variable but that is not affected by the dependent variable.

influential data point. A data point whose deletion from a regression sample causes one or more estimated regression parameters to change substantially.

instrumental variable. A variable that is both highly correlated with the covariate that is believed to be endogenous and at the same time not directly affected by the dependent variable.

interaction variable. The product of two explanatory variables in a regression model; used in a particular form of nonlinear model.

intercept. The value of the dependent variable when each of the explanatory variables takes on the value of 0 in a regression equation.

least squares (or ordinary least squares). A common method for estimating regression parameters. Least squares minimizes the sum of the squared differences between the actual values of the dependent variable and the values predicted by the regression equation.

linear regression model. A regression model in which the effect of a change in each of the explanatory variables on the dependent variable is the same, no matter what the values of those explanatory variables.

logit model. A discrete dependent-variable regression model in which the estimated coefficients measure the effect of a change in the covariates on the logarithm of the odds that a particular choice will be made or an event will occur.

maximum likelihood estimation. An estimation method that determines values for the parameters of the regression model that maximizes the likelihood that the process described by the model produced the sample data that were observed.

mean; expected value. An average of the outcomes associated with a probability distribution, where the outcomes are weighted by the probability that each will occur.

mean squared error (MSE). The estimated variance of the regression error, calculated as the average of the sum of the squares of the regression residuals.

model. A mathematical representation of an actual situation.

multicollinearity. When two or more variables are highly correlated in a multiple regression analysis. Substantial multicollinearity can cause regression parameters to be estimated imprecisely, as reflected in relatively high standard errors.

multiple regression analysis. A statistical tool for understanding the relationship between two or more variables.

nonlinear regression model. A model having the property that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change.

normal distribution. A bell-shaped probability distribution having the property that about 95% of the distribution lies within two standard deviations of the mean.

null hypothesis. In regression analysis, the null hypothesis states that the results observed in a study with respect to a particular variable are no different from what might have occurred by chance, independent of the effect of that variable. See hypothesis test.

one-tailed test. A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, but not both.

outlier. A data point that is more than some appropriate distance from a regression line that is estimated using all the other data points in the sample.

***p*-value.** The significance level in a statistical test; the probability of getting a test statistic as extreme or more extreme than the observed value. The larger the *p*-value, the more likely that the null hypothesis is valid.

parameter. A numerical characteristic of a population or a model.

perfect collinearity. When two or more explanatory variables are correlated perfectly.

population. All the units of interest to the researcher; also, universe.

practical significance. Substantive importance. Statistical significance does not ensure practical significance because, with large samples, small differences can be statistically significant.

probability distribution. The process that generates the values of a random variable. A probability distribution lists all possible outcomes and the probability that each will occur.

probability sampling. A process by which a sample of a population is chosen so that each unit of observation has a known probability of being selected.

probit model. A discrete dependent-variable regression model in which the estimated coefficients measure the effect of a change in a covariate on the probability that a particular choice will be made or an event will occur.

quasi-experiment (or natural experiment). A naturally occurring instance of observable phenomena that yield data that approximate a controlled experiment.

***R*-squared (R^2).** A statistic that measures the percentage of the variation in the dependent variable that is accounted for by all of the explanatory variables in a regression model. *R*-squared is the most used measure of goodness of fit of a regression model.

random-effects regression model. A regression model that views individual-specific constant terms as being randomly distributed across the individual units of observation.

random error term. A term in a regression model that reflects random error (sampling error) that is the result of chance. Consequently, the result obtained

in the sample differs from the result that would be obtained if the entire population were studied.

randomized controlled trial (RCT). A methodology in which the analyst randomly divides potential subjects into two groups: the treatment group, who receive an intervention, and the control group, who do not.

regression coefficient. The estimate of a population parameter obtained from a regression equation that is based on a particular sample; also, regression parameter.

regression residual. The difference between the actual value of a dependent variable and the value predicted by the regression equation.

robust estimation. An alternative to least-squares estimation that is less sensitive to outliers.

robust. When a statistic or procedure does not change much when data or assumptions are slightly modified.

sample. A selection of data chosen for a study; a subset of a population.

sampling error. A measure of the difference between the sample estimate of a parameter and the population parameter.

scatterplot. A graph showing the relationship between two variables in a study; each dot represents one subject. One variable is plotted along the horizontal axis; the other variable is plotted along the vertical axis.

serial correlation. The correlation of the values of regression errors over time.

simultaneity. When a covariate in a regression model affects the dependent variable and is also affected by the dependent variable.

slope. The change in the dependent variable associated with a one-unit change in an explanatory variable in a linear regression model.

spurious correlation. When two variables are correlated, but one is not the cause of the other.

standard deviation. The square root of the variance of a random variable. The variance is a measure of the spread of a probability distribution about its mean; it is calculated as a weighted average of the squares of the deviations of the outcomes of a random variable from its mean.

standard error of forecast (SEF). An estimate of the standard deviation of the forecast error; it is based on forecasts made within a sample in which the values of the explanatory variables are known with certainty.

standard error of the coefficient; standard error (SE). A measure of the variation of a parameter estimate about the true parameter. The standard error is a standard deviation that is calculated from the probability distribution of estimated parameters.

standard error of the regression (SER). An estimate of the standard deviation of the regression error; it is calculated as the square root of the average of the squares of the residuals associated with a particular multiple regression analysis.

statistical significance. A test used to evaluate the degree of association between a dependent variable and one or more explanatory variables. If the calculated *p*-value is smaller than 5%, the result is said to be statistically significant (at the 5% level). If *p* is greater than 5%, the result is statistically insignificant (at the 5% level).

***t*-statistic.** A test statistic that describes how far an estimate of a parameter is from its hypothesized value (i.e., given a null hypothesis). If a *t*-statistic is sufficiently large (in absolute magnitude), an expert can reject the null hypothesis.

***t*-test.** A test of the null hypothesis that a regression parameter takes on a particular value, usually 0. The test is based on the *t*-statistic.

time-series analysis. A type of multiple regression analysis in which each data point is associated with a particular unit of observation (e.g., an individual or a firm) measured at different points in time.

two-tailed test. A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, or both.

variable. Any attribute, phenomenon, condition, or event that can have two or more values.

variable of interest. The explanatory variable that is the focal point of a particular study or legal issue.

References on Multiple Regression

Orley Ashenfelter, Theodore Eisenberg & Stewart J. Schwab, *Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes*, 24 J. Legal Stud. 257 (1995).

Jonathan A. Baker & Daniel L. Rubinfeld, *Empirical Methods in Antitrust: Review and Critique*, 1 Am. L. & Econ. Rev. 386 (1999).

Gerald V. Barrett & Donna M. Sansonetti, *Issues Concerning the Use of Regression Analysis in Salary Discrimination Cases*, 41 Personnel Psych. 503 (2006).

Leo Breinman, *Random Forests*, 45 Machine Learning 5 (2001).

Thomas J. Campbell, *Regression Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 Stan. L. Rev. 1299 (1984).

Catherine Connolly, *The Use of Multiple Regression Analysis in Employment Discrimination Cases*, 10 Population Res. & Pol'y Rev. 117 (1991).

Arthur P. Dempster, *Employment Discrimination and Statistical Science*, 3 Stat. Sci. 149 (1988).

The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989).

Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 Colum. L. Rev. 737 (1980).

Michael O. Finkelstein & Hans Levenbach, *Regression Estimates of Damages in Price-Fixing Cases*, 46 Law & Contemp. Probs. 145.

Franklin M. Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. Am. Stat. Ass'n 277 (1986).

Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 Colum. L. Rev. 702 (1980).

Joseph L. Gastwirth, *Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables*, 33 Jurimetrics J. 19 (1992).

Peter Kennedy, A Guide to Econometrics (6th ed. 2019).

Note, *Beyond the Prima Facie Case in Employment Discrimination Law: Statistical Proof and Rebuttal*, 89 Harv. L. Rev. 387 (1975).

Daniel L. Rubinfeld, *Econometrics in the Courtroom*, 85 Colum. L. Rev. 1048 (1985).

Daniel L. Rubinfeld & Peter O. Steiner, *Quantitative Methods in Antitrust Litigation*, 46 Law & Contemp. Probs. 69 (1983).

Daniel L. Rubinfeld, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991).

James H. Stock & Mark W. Watson, Introduction to Econometrics (4th ed., 2018).

References on Causal Analysis

Joseph G. Altonji, Todd E. Elder & Christopher R. Taber, *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools*, 113 J. Pol. Econ. 151 (2005).

Isaiah Andrews & Timothy B. Armstrong, *Unbiased Instrumental Variables Estimation Under Known First-Stage Sign*, 8 Quantitative Econ. 479 (2017).

Isaiah Andrews, James H. Stock & Liyang Sun, *Weak Instruments in IV Regression: Theory and Practice*, 11 Ann. Rev. Econ. 727 (2019).

Joshua D. Angrist, *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, 80 Am. Econ. Rev. 313 (1990).

Orley Ashenfelter, *Estimating the Effect of Training Programs on Earnings*, 60 Rev. Econ. & Stat. 47 (1978).

Andrew C. Baker, David F. Larcker & Charles C. Y. Wang, *How Much Should We Trust Staggered Difference-in-Differences Estimates?*, 144 J. Fin. Econ. 370 (2022).

John Bound, David A. Jaeger & Regina M. Baker, *Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak*, 90 J. Am. Stat. Ass'n 443 (1995).

David Card, *The Causal Effect of Education on Earnings, in* 3 Handbook of Labor Economics, 1801 (Orley Ashenfelter & David Card eds., 1999).

David Card & Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, 84 Am. Econ. Rev. 772 (1994).

Clément de Chaisemartin & Xavier D'Haultfœuille, *Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects*, 110 Am. Econ. R. 2964 (2020).

David Chan, Matthew Gentzkow & Chuan Yu, *Selection with Variation in Diagnostic Skill*, 137 Q. J. Econ. 729 (2022).

Maria De Paola & Vincenzo Scoppa, *The Effects of Managerial Turnover: Evidence from Coach Dismissals in Italian Soccer Teams*, 12 J. Sports Econ. 132 (2012).

Will Dobbie, Jacob Goldin & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*. 108 Am. Econ. Rev. 201 (2012).

Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

Andrew Goodman-Bacon, *Difference-in-Differences with Variation in Treatment Timing*, 225 J. Econometrics 254 (2021).

Justine S. Hastings, *Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California*, 94 Am. Econ. Rev. 317–28 (2004).

Guido W. Imbens & Joshua D. Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 Econometrica 467–75 (1994).

Louis S. Jacobson, Robert J. Lalonde & Daniel G. Sullivan, *Earnings Losses of Displaced Workers*, 83 Am. Econ. Rev. 685–709 (1993).

Ginger Zhe Jin & Phillip Leslie, *The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards*, 118 Q. J. Econ. 409–51 (2003).

Peter E. Kennedy, *Sinning in the Basement: What are the Rules? The Ten Commandments of Applied Econometrics*, J. Econ. Surveys 578 (2002).

Edward E. Leamer, Specification Searches: Ad Hoc Inference with Nonexperimental Data (1978).

Steven D. Levitt, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime*, 87 Am. Econ. Rev. 270–90 (1997).

Charles E. Loeffler, *Pre-Imprisonment Employment Drops: Another Instance of the Ashenfelter Dip?*, 108 J. Crim. Law & Criminology 815–38 (2018).

Winston Lin, *Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique*, 7 Annals Applied Stats. 295–318 (2013).

Justin McCrary, *The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police*, 97 Am. Econ. Rev. 318–53 (2007).

Tracey L. Meares, *The Law and Social Science of Stop and Frisk*, 10 Ann. Rev. L. & Soc. Sci. 335–52 (2014).

Ashley Miller, *Principal turnover and student achievement*, 6 Econ. Educ. Rev. 60 (2013).

Emily Oster, *Unobservable Selection and Coefficient Stability: Theory and Evidence*, 37 J. of Bus. & Econ. Stats. 187–204 (2019).

Peter H. Rossi, Richard A. Berk & Kenneth J. Lenihan, Money, Work and Crime: Experimental Evidence (1980).

Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. Educ. Psych. 688–701 (1974).

Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 Am. Socio. Rev. 261–72 (1984).

James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518–29 (2002).

James H. Stock & Mark W. Watson, Introduction to Econometrics (4th ed. 2019).

Jeffrey M. Wooldridge, Econometric Analysis of Cross Section and Panel Data (2001).

Reference Guide on Survey Research

SHARI SEIDMAN DIAMOND, MATTHEW KUGLER, AND JAMES N. DRUCKMAN

Shari Seidman Diamond, J.D., Ph.D., is the Howard J. Trienens Professor of Law and Professor of Psychology at Northwestern University and a Research Professor at the American Bar Foundation, Chicago.

Matthew Kugler, J.D., Ph.D., is Professor of Law at Northwestern University.

James N. Druckman, Ph.D., is Martin Brewer Anderson Professor of Political Science at the University of Rochester.

CONTENTS

Introduction

Sample surveys gather responses from a subset of a population to draw inferences about the population as a whole. They are used to describe or enumerate beliefs, attitudes, or behaviors of persons or other social units.[1] Surveys typically are offered in legal proceedings to establish or refute claims about the characteristics of those individuals or social units.[2] We focus here primarily on sample surveys with individuals reporting about themselves (e.g., their own beliefs) or their organizations (e.g., the company where they are employed). Such surveys must deal not only with issues of population definition, sampling, and measurement common to all surveys, but also with the specialized issues that arise in obtaining information from human respondents.

In principle, a survey can count or measure every member of the relevant population. A survey that does this full count is sometimes called a census. In practice, however, most surveys typically count or measure only a portion of the individuals or other units that the survey intends to describe. In either case, the goal is to provide information on the relevant population. Sample surveys can be carried out using probability or nonprobability sampling techniques. Although probability sampling offers important advantages over nonprobability sampling,[3] various forms of nonprobability sampling are in wide use. Thus, in this reference guide, we discuss both probability samples and nonprobability samples, including their strengths and weaknesses for achieving various purposes.

As a method of data collection, surveys have several crucial potential advantages over less systematic approaches.[4] When a survey is properly designed,

1. Sample surveys conducted by social scientists "consist of (relatively) systematic, (mostly) standardized approaches to collecting information on individuals, households, organizations, or larger organized entities through questioning systematically identified samples." James D. Wright & Peter V. Marsden, *Survey Research and Social Science: History, Current Practice, and Future Prospects*, *in* Handbook of Survey Research 1, 3 (James D. Wright & Peter V. Marsden eds., 2d ed. 2010).

2. *See, e.g.*, Sanderson Farms, Inc. v. Tyson Foods, Inc., 547 F. Supp. 2d 491 (D. Md. 2008); SMS Sys. Maint. Servs. v. Digital Equip. Corp. 188 F.3d 11 22, 22–23 (1st Cir. 1999). For other examples, see *infra* notes 10–26 and accompanying text.

3. *See* section titled "The Sample as a Reflection of the Relevant Characteristics of the Population" below.

4. This does not mean that surveys can be relied on to address all questions. For example, if survey respondents had been asked in the days before the attacks of 9/11 to predict whether they would volunteer for military service if Washington, D.C., were to be bombed, their answers may not have provided accurate predictions. Although respondents might have willingly answered the question, their assessment of what they would actually do in response to an attack simply may have been inaccurate. Even the option of a "do not know" choice would not have prevented an error in prediction if they believed they could accurately predict what they would do. Thus, although such a survey would have been suitable for assessing the *predictions* of respondents, it might have provided a very inaccurate estimate of what an actual response to the attack would be. If a survey respondent has limited experience (e.g., a child predicting what she will do as an adult) or the hypothetical is far removed from reality or imaginable reality, the survey is unlikely to provide accurate predictions.

executed, and described, it (1) efficiently presents the responses of a group of individuals or other units (e.g., organizations) and (2) permits an assessment of the extent to which the measured responses of the individuals or other units are likely to adequately represent a relevant population of individuals or other units. All questions asked of respondents and all other measuring devices used (e.g., criteria for selecting eligible respondents) can be examined by the court and the opposing party for objectivity, clarity, and relevance, and all answers or other measures obtained can be analyzed for completeness and consistency.

So that the court and the opposing party can closely scrutinize the survey, the party offering the survey as evidence should describe in detail the design, execution, and analysis of the survey, including (1) a description of the population from which the sample was selected, demonstrating that it was a relevant population for the question at hand; (2) a description of how the sample was drawn and an explanation for why that sample design was appropriate; (3) a report on response rate and the ability of the sample to represent the target population; (4) evidence that respondents were attentive and honest in answering the questions on the survey; and (5) an evaluation of any sources of potential bias in respondents' answers or in the ability of the results from the sample to generalize to the relevant population.

The material covered in this reference guide is intended to assist judges in identifying, narrowing, and addressing issues bearing on the adequacy of surveys either offered as evidence or proposed as a method for developing information. Questions about a survey can be (1) raised from the bench during a pretrial proceeding to determine the admissibility of the survey evidence; (2) presented to the contending experts before trial for their joint identification of disputed and undisputed issues; (3) presented to counsel with the expectation that the issues will be addressed during the examination of the experts at trial; or (4) raised in bench trials to help the judge evaluate what weight, if any, the survey should be given.[5]

All sample surveys should address the issues concerning purpose and design (see section titled "Purpose and Design of the Survey" below), population definition and sampling (see section titled "Population Definition and Sampling" below), and disclosure and reporting (see section titled "Disclosure and Reporting" below). All questionnaire and interview surveys raise methodological issues involving survey questions and structure (see section titled "Survey Questions and Structure" below) and confidentiality (see section titled "Disclosure and Reporting" below). Interview surveys introduce additional issues, such as accuracy of data entry (see section titled "Accuracy of Data Entry" below) and interviewer training and qualifications (see

5. Lanham Act cases involving trademark infringement or deceptive advertising frequently require expedited hearings that request injunctive relief, and so judges may need to be more familiar with survey methodology when considering the weight to accord a survey in these cases than when presiding over cases being submitted to a jury. Even in a case being decided by a jury, however, the court must be prepared to evaluate the methodology of the survey evidence in order to rule on admissibility. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 589 (1993); Fed. R. Evid. 702.

section titled "Surveys Involving Interviewers" below). And online surveys raise special issues and questions (see section titled "Internet Surveys" below).

Use of Surveys in Court

Sixty years ago, the question of whether surveys were acceptable evidence was unsettled.[6] Early doubts about the admissibility of surveys centered on their use of sampling and their status as hearsay evidence. Federal Rule of Evidence 703 settled both matters for surveys by redirecting attention to the "validity of the techniques employed."[7] The inquiry under Rule 703 focuses on whether facts or data are "of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject."[8]

Because the survey method provides an economical and systematic way to gather information and draw inferences about a large number of individuals or other units, surveys are used widely in business, government, administrative settings, and judicial proceedings.[9] Both federal and state courts have accepted survey evidence on a variety of issues. Our review of cases citing survey evidence over the ten-year period between 2012 and 2022 revealed cases involving legal issues in administrative law, copyright, criminal law, deceptive advertising, discrimination, employment, patent, trademark, and unfair competition, among others.

Some of these cases cited surveys conducted specifically for litigation, and some cited surveys not prepared for litigation. The topics of both litigation and nonlitigation surveys are diverse.[10] Surveys appear even in the most routine of

6. Hans Zeisel, *The Uniqueness of Survey Evidence*, 45 Cornell L.Q. 322, 345 (1960).

7. Fed. R. Evid. 703 advisory committee note. This focus on the adequacy of the methodology used in conducting and analyzing results from a survey is also consistent with the Supreme Court's discussion of admissible scientific evidence in *Daubert*, 509 U.S. 579; *see also* General Elec. Co. v. Joiner, 522 U.S. 136, 147 (1997).

8. Fed. R. Evid. 703 advisory committee note.

9. Some sample surveys are so well accepted that they may not even be recognized as surveys. For example, some U.S. Census Bureau data are based on sample surveys, including the widely relied-upon American Community Survey, https://perma.cc/XLY8-R4YE. Similarly, the Standard Table of Mortality, which is accepted as proof of the average life expectancy of an individual of a particular age and gender, is based on survey data. Surveys conducted by federal agencies are generally of high quality. Their demographic statistics are often used as benchmarks for nongovernmental research.

10. *See, e.g.*, Kittle-Aikeley v. Claycomb, 807 F.3d 913, 926 (8th Cir. 2015) (survey to assess the seriousness of the drug abuse problem in schoolchildren to help justify the school's decision to drug test its community college student body); Miller v. Bonta, 542 F. Supp. 3d 1009, 1021 (S.D. Cal. 2021), *vacated and remanded*, No. 21–55608, 2022 WL 3095986 (9th Cir. Aug. 1, 2022) (in a Second Amendment case, survey used to note that many people in California own firearms); Missouri v. Biden, 576 F. Supp. 3d 622, 634 (E.D. Mo. 2021) (survey used to predict workers' response to a vaccine mandate for federal contractors); United States v. Cloud, No. 1:19-cr-02032-SMJ-1, 2021 U.S. Dist. LEXIS 235350, at *13 (E.D. Wash. Dec. 8, 2021) (survey used to determine whether a jury pool was sufficiently biased to warrant a change of venue); Lord & Taylor LLC v. Zim Integrated Shipping Servs., Ltd., 108 F. Supp. 3d 197, 227 (S.D.N.Y. 2015) (survey used to note that 79% of

cases. Vocational experts testifying in Social Security Administration disability hearings regularly rely on a variety of surveys conducted by the Bureau of Labor Statistics, specifically the Occupational Requirements Survey,[11] the Occupational Employment Survey,[12] and the National Compensation Survey,[13] to support their opinions on whether a person would be able to find gainful employment given their documented disabilities. One might think that reliance on these surveys is unexceptional, but federal courts have sometimes struggled to determine whether a given expert's use of survey data is scientifically valid.[14]

Employment and discrimination law cases sometimes incorporate survey evidence as well. These surveys may be conducted as part of a company's normal business operations and may be relevant to the performance of a particular employee or the feelings of a subgroup of employees.[15] Other times the surveys may be conducted by plaintiffs to provide evidence of wage and working conditions.[16]

In cases in which courts set attorneys' fees, judges are charged with determining reasonable fees.[17] Surveys frequently provide courts with evidence of the prevailing market rates.[18]

coastal residents thought the impact of Hurricane Sandy was worse than expected, informing whether it should be treated as an Act of God for contract purposes).

11. Roberta G. v. Kijakazi, No. CV 20-07796-DFM, 2021 U.S. Dist. LEXIS 179074, at *6–7 (C.D. Cal. Sept. 20, 2021).

12. Sok v. Kijakazi, No. 20-cv-489-wmc, 2021 U.S. Dist. LEXIS 183024, at *12 (W.D. Wis. Sept. 24, 2021); Dawn L.C. v. Comm'r of Soc. Sec., No. 3:20-cv-00626-GCS, 2021 U.S. Dist. LEXIS 191829, at *19 (S.D. Ill. Sept. 24, 2021); Stephanie D. v. Comm'r of Soc. Sec., No. 3:20-CV-0768 (ML), 2021 U.S. Dist. LEXIS 168701, at *8 (N.D.N.Y. Sept. 7, 2021).

13. Missey v. Saul, No. 4:20-CV-701-ERW, 2021 U.S. Dist. LEXIS 120435, at *14 (E.D. Mo. June 29, 2021).

14. Alaura v. Colvin, 797 F.3d 503, 507–08 (7th Cir. 2015) (describing one apparently common practice in apportioning job availability numbers extracted from the Occupational Employment Survey as "preposterous"). *See also Dawn L.C.*, 2021 U.S. Dist. LEXIS 191829, at *17 (describing the subsequent debate among the district courts).

15. Spivey v. Mohawk ESV, Inc., No. 7:19-cv-00670, 2021 U.S. Dist. LEXIS 111905, at *3 (W.D. Va. June 15, 2021) (use of a manager's staff approval ratings in an age discrimination case); Griffin v. Shelby Residential & Vocational Servs., No. 2:18-cv-2665, 2021 U.S. Dist. LEXIS 111866, at *24 (W.D. Tenn. June 15, 2021) ("The Court finds that the 2016 QA surveys are the most probative and objective evidence of Defendant's rationale for termination."); Milioto-Maruca v. Lauren, No. 3:11-CV-2120, 2012 U.S. Dist. LEXIS 173945, at *5 (M.D. Pa. Dec. 7, 2012) (work climate survey was conducted at the stores managed by the plaintiff in response to subordinate complaints).

16. Medlock v. Taco Bell Corp., No. 1:07-cv-01314-SAB, 2015 U.S. Dist. LEXIS 165235 (E.D. Cal. Dec. 9, 2015).

17. Courts are directed to calculate a lodestar figure by multiplying the number of hours reasonably expended on the litigation times a reasonable hourly rate and then adjust upward or downward based on particular features of the work involved. Jones v. George Fox Univ., No. 3:19-cv-0005-JR, 2022 U.S. Dist. LEXIS 162869, at *3 (D. Or. Sept. 9, 2022) (citing Blum v. Stenson, 465 U.S. 886, 888 (1984)).

18. *See, e.g.*, *Jones*, 2022 U.S. Dist. LEXIS 162869, at *8–9 (a survey of attorneys was conducted by the Portland State University Survey Research Lab for the Oregon State Bar; the fee

Surveys also are common in the areas of trademark and false-advertising law. Survey evidence has been used to establish whether a proposed mark is a generic term,[19] assess whether a mark has secondary meaning or achieved sufficient fame for a dilution claim,[20] evaluate whether a defendant's product presents a likelihood of confusion with a plaintiff's mark,[21] and probe how consumers would have interpreted an allegedly misleading advertisement.[22]

Surveys have additionally been conducted in false-advertising cases to assess the value of the deceptive claim to consumers[23] and in patent cases to assess the value of product features.[24] The use of conjoint analysis, a relatively new method of making such value attributions for litigation, is discussed below.[25]

On occasion, courts ruling on the admissibility of scientific claims have examined surveys of scientific experts to assess the extent to which a theory or

award adopted by the court used the 95th percentile rate in light of the attorney's prior experience before admission to the bar).

19. *See, e.g.*, Snyder's Lance, Inc. v. Frito-Lay N. Am., Inc., 542 F. Supp. 3d 371, 397–99, 401–02 (W.D.N.C. 2021) (contrasting the results of two surveys considering "Pretzel Crisps"); Primary Children's Med. Ctr. Found. v. Scentsy, Inc., No. 2:11-cv-1141-TC, 2012 U.S. Dist. LEXIS 86318, at *12 (D. Utah June 20, 2012) (finding important the results of a Teflon survey analyzing perceptions of "Festival of Trees").

20. Warner Bros. Ent. v. Glob. Asylum, Inc., No. CV 12–9547 PSG (CWx), 2012 U.S. Dist. LEXIS 185695, at *11 (C.D. Cal. Dec. 10, 2012) ("The survey results show[] that nearly 50 percent of respondents associated the term 'Hobbit' with" a single source); MZ Wallace Inc. v. Fuller, No. 18cv2265(DLC), 2018 U.S. Dist. LEXIS 214754, at *32–35 (S.D.N.Y. Dec. 20, 2018) (finding helpful a study showing a lack of secondary meaning in the plaintiff's alleged mark); ProFoot, Inc. v. MSD Consumer Care, Inc., No. 11–7079, 2012 U.S. Dist. LEXIS 83427, at *25–26 (D.N.J. June 14, 2012) (insufficient awareness to support a claim of fame).

21. Under Armour, Inc. v. Battle Fashions, Inc., No. 5T9-CV-297-BO, 2021 U.S. Dist. LEXIS 114292, at *5–6 (E.D.N.C. June 18, 2021) (admitting a survey evaluating whether the defendant's use of "I can do all things" infringes on the plaintiff's mark); Nat'l Fin. Partners Corp. v. Paycom Software, Inc., No. 14 C 7424, 2015 U.S. Dist. LEXIS 74700, at *32–33 (N.D. Ill. June 10, 2015) (finding probative one of the two Squirt-style studies conducted to assess likelihood of confusion).

22. Naimi v. Starbucks Corp., 798 F. App'x 67, 69 (9th Cir. 2019) (plaintiff's survey sufficient to establish implied representation at the motion to dismiss stage); Benson v. Newell Brands, Inc., No. 19 C 6836, 2021 U.S. Dist. LEXIS 220986, at *19–22 (N.D. Ill. Nov. 16, 2021); *In re* Elysium Health-ChromaDex Litig., No. 17-cv-7394 (LJL), 2022 U.S. Dist. LEXIS 25063, at *5–36 (S.D.N.Y. Feb. 11, 2022) (contrasting the results of two experts' false advertising studies).

23. *In re* Dial Complete Mktg. & Sales Pracs. Litig., 320 F.R.D. 326 (D.N.H. 2017) (alleged damages, claiming that the soap did not kill 99.9% of germs, but some smaller percentage).

24. Apple, Inc. v. Samsung Elecs. Co., No. 11-CV-01846-LHK, 2012 U.S. Dist. LEXIS 90877, at *37–38, *42–43 (N.D. Cal. June 29, 2012); SimpleAir, Inc. v. Google Inc., No. 2:14-CV-11, 2015 U.S. Dist. LEXIS 135915, at *6–7 (E.D. Tex. Oct. 5, 2015); TV Interactive Data Corp. v. Sony Corp., 929 F. Supp. 2d 1006, 1020–22 (N.D. Cal. 2013); Microsoft Corp. v. Motorola, Inc., 904 F. Supp. 2d 1109, 1119–20 (W.D. Wash. 2012).

25. *See infra* pp. 722–25.

technique has received widespread acceptance.[26] Such surveys can be valuable in assisting the court, but they must reflect the views of a representative group of recognized experts who have responded to questions about the relevant issue.

In addition, survey methodology has been used creatively to assist federal courts in managing mass torts litigation. For instance, faced with the prospect of conducting discovery concerning 10,000 plaintiffs, the plaintiffs and defendants in *Wilhoite v. Olin Corp.*[27] jointly drafted a discovery survey that was administered in person by neutral third parties, thus replacing interrogatories and depositions. It resulted in substantial savings in both time and cost. Greater use of this approach would be beneficial to all parties.

A Comparison of Survey Evidence and Individual Testimony

To illustrate the value of a survey, it is useful to compare the information that can be obtained from a competently done survey with the information obtained by other means. A survey is presented by a survey expert, who testifies about the responses of a substantial number of individuals who have been selected according to an explicit sampling plan and asked the same set of questions. In contrast, a party using a nonsurvey method generally identifies several witnesses who testify about their own characteristics, experiences, or impressions. Although the party has no obligation to select these witnesses in any particular way or to report on how they were chosen, the party is not likely to select witnesses whose attitudes or beliefs conflict with the party's interests. The witnesses who testify are aware of the parties involved in the case and have discussed the case with the party before testifying.

Although surveys are not the only means of demonstrating particular facts, presenting the results of a well-done survey through the testimony of an expert is an efficient way to inform the trier of fact about a large group of potential witnesses. In some cases, courts have described surveys as the most direct form of

26. For instance, courts determined that the polygraph test has failed to achieve general acceptance in the scientific community based on the inconsistent reactions revealed in several surveys. *See* United States v. Scheffer, 523 U.S. 303, 309–10 (1998); United States v. Bishop, 64 F. Supp. 2d 1149 (D. Utah 1999); United States v. Varoudakis, No. 97-10158-RGS, 1998 WL 151238 (D. Mass. Mar. 27, 1998); State v. Shively, 999 P.2d 952, *aff'd*, 999 P.2d 259 (Kan. 2000); Lee v. Martinez, 96 P.3d 291, 304–06 (N.M. 2004). In contrast, an eyewitness identification expert was permitted to testify about scientific studies of factors affecting the perceptual ability and memory of eyewitnesses based in part on survey evidence showing general acceptance of those findings by experts within the field. *See* People v. Williams, 830 N.Y.S.2d 452, 465–66 (N.Y. Sup. Ct. 2006).

27. Wilhoite v. Olin Corp., No. CV-83-C-5021-NE (N.D. Ala. filed Jan. 11, 1983). The case ultimately settled before trial. *See* Francis E. McGovern & E. Allan Lind, *The Discovery Survey*, 51 Law & Contemp. Probs. 41, 49 (1988).

evidence that can be offered.[28] Indeed, several courts have drawn negative inferences from the absence of a survey, taking the position that failure to undertake a survey may strongly suggest that a properly done survey would not support the plaintiff's position.[29]

Purpose and Design of the Survey

Addressing Relevant Questions

Key to evaluating any survey is assessing whether it is relevant to the disputed issues. Surveys conducted in the normal course of business and not in anticipation of, or in response to, litigation may have a lesser risk of bias in favor of the interests of the party conducting the survey, but such surveys may ask irrelevant questions, lack important quality controls, or be conducted on inappropriate populations.[30] In contrast, surveys conducted for litigation are more likely to be designed to address the legally relevant issues in the case (e.g., to estimate damages in an antitrust suit or to assess consumer confusion in a trademark case), but may have a greater risk of bias since they are typically solicited by one of the parties to aid in the case. Thus, the content and execution of a survey must be scrutinized whether or not the survey was explicitly designed to provide relevant data on the issue before the court.[31]

28. *See, e.g.*, Morrison Ent. Grp. v. Nintendo of Am., 56 F. App'x 782, 785 (9th Cir. 2003); Monster, Inc. v. Dolby Lab'ys Licensing Corp., 920 F. Supp. 2d 1066, 1072 (N.D. Cal. 2013); Heaven Hill Distilleries, Inc. v. Log Still Distilling, LLC, No. 3:21-cv-190-BJB-CHL, 2021 U.S. Dist. LEXIS 240373 (W.D. Ky. Dec. 16, 2021) (survey is the "most persuasive evidence" of consumer recognition).

29. Ortho Pharm. Corp. v. Cosprophar, Inc., 32 F.3d 690, 695 (2d Cir. 1994); Henri's Food Prods. Co. v. Kraft, Inc., 717 F.2d 352, 357–58 (7th Cir. 1983); Medici Classics Prods. LLC v. Medici Grp. LLC, 590 F. Supp. 2d 548, 556 (S.D.N.Y. 2008); Citigroup v. City Holding Co., No. 99 Civ. 10115 (RWS), 2003 U.S. Dist. LEXIS 1845 (S.D.N.Y. Feb. 10, 2003); Chum Ltd. v. Lisowski, 198 F. Supp. 2d 530 (S.D.N.Y. 2002); Cairns v. Franklin Mint Co., 24 F. Supp. 2d 1013, 1041 (C.D. Cal. 1998) ("[A] plaintiff's failure to conduct a survey, assuming it has the financial resources to do so, may lead to an inference that the results of such a survey would be unfavorable.").

30. In *Craig v. Boren*, 429 U.S. 190 (1976), the state unsuccessfully attempted to use its annual roadside survey of the blood alcohol level, drinking habits, and preferences of drivers to justify prohibiting the sale of 3.2% beer to males under the age of 21 and to females under the age of 18. The data were biased because it was likely that the male would be driving if both the male and female occupants of the car had been drinking. As pointed out in 2 Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy: Tort Law, Evidence, and Health* 527 (1988), the roadside survey would have provided more relevant data if all occupants of the cars had been included in the survey (and if the type and amount of alcohol most recently consumed had been requested so that the consumption of 3.2% beer could have been isolated).

31. *See* Merisant Co. v. McNeil Nutritionals, LLC, 242 F.R.D. 315 (E.D. Pa. 2007).

Role of Attorneys in Survey Design and Administration

An early handbook for judges recommended that survey interviews be "conducted independently of the attorneys in the case."[32] Some courts interpreted this to mean that any evidence of attorney participation is objectionable.[33] A better interpretation is that the attorney should not take part in survey implementation or interact with survey participants.[34] However, some attorney involvement in the survey design is often necessary to ensure that relevant questions are directed to a relevant population,[35] particularly if the survey expert is not an expert in the substantive area of law under dispute. Federal Rule of Civil Procedure 26(4)(b) does not allow an inquiry into the nature of communications between attorneys and experts, and so the role of attorneys in constructing surveys may not always be fully apparent. The key issues for the trier of fact are the design of the survey, the objectivity and relevance of the questions on the survey, the appropriateness of the population used to guide sample selection, and the method of sample selection. These aspects of the survey are visible to the trier of fact and can be judged on their quality, irrespective of who suggested them. In contrast, the survey administration itself, whether online or in an interview, may not be directly visible. Any potential bias is minimized by having interviewers and respondents blind to the purpose and sponsorship of the survey and by excluding attorneys from any part in administering questionnaires, conducting interviews, tabulating results, and interpreting the data.[36]

32. Judicial Conference of the United States, Handbook of Recommended Procedures for the Trial of Protracted Cases, 25 F.R.D. 351, 429 (1960).

33. *See, e.g.*, Boehringer Ingelheim G.m.b.H. v. Pharmadyne Lab'ys, 532 F. Supp. 1040, 1058 (D.N.J. 1980).

34. Upjohn Co. v. Am. Home Prods. Corp., No. 1-95-CV-237, 1996 U.S. Dist. LEXIS 8049, at *42 (W.D. Mich. Apr. 5, 1996) (objection that "counsel reviewed the design of the survey carries little force with this Court because [opposing party] has not identified any flaw in the survey that might be attributed to counsel's assistance"). For cases in which attorney participation was linked to significant flaws in the survey design or execution, see *Hurt v. Commerce Energy, Inc.*, No. 1:12-CV-00758, 2015 U.S. Dist. LEXIS 10566 (N.D. Ohio Jan. 29, 2015); *Johnson v. Big Lots Stores, Inc.,* No. 04–321, 2008 U.S. Dist. LEXIS 35316, at *52–53 (E.D. La. Apr. 29, 2008); *United States v. Southern Indiana Gas & Electric Co.*, 258 F. Supp. 2d 884, 894 (S.D. Ind. 2003); and *Gibson v. County of Riverside*, 181 F. Supp. 2d 1057, 1069 (C.D. Cal. 2002).

35. *See* 6 J. Thomas McCarthy, McCarthy on Trademarks and Unfair Competition § 32:166 (5th ed. 2021). *See also* Jerre B. Swann, *A History of the Evolution of Likelihood of Confusion Methodologies*, 113 Trademark Rep. 723 (2023) (generally on the importance of context in determining survey design methodology).

36. *Gibson,* 181 F. Supp. 2d at 1068.

Skill and Experience of the Experts Who Designed, Conducted, or Analyzed the Survey

Experts prepared to design, conduct, and analyze a survey generally should have graduate training in psychology (especially social, cognitive, or consumer psychology), sociology, political science, marketing, communication sciences, statistics, or a related discipline; that training should include courses in survey research methods, sampling, measurement, interviewing, and statistics. In some cases, professional experience in teaching or conducting and publishing survey research may provide the requisite background. In all cases, the expert must demonstrate an understanding of best practices in survey methodology, including sampling,[37] instrument design (questionnaire and interview construction), and statistical analysis.[38] Publication in peer-reviewed journals, authored books, fellowship status in professional organizations, faculty appointments, consulting experience, research grants, and membership on scientific advisory panels for government agencies or private foundations are indications of a professional's area and level of expertise. In addition, some surveys involving highly technical subject matter or specific (sub)populations may require experts to have some further specialized knowledge. Under these conditions, the survey expert also should be able to demonstrate sufficient familiarity with the topic and population (or assistance from an individual on the research team with suitable expertise) to design a survey instrument that will communicate clearly with relevant respondents.

Skill and Experience of the Experts Who Will Testify About Surveys Conducted by Others

Parties often call on an expert to testify about a survey conducted by someone else (e.g., by one of the parties to the suit, or by another entity when the survey was not conducted specifically for the case). The secondary expert's role may be to offer support for a survey commissioned by the party who calls the expert, to critique a survey presented by the opposing party, or to introduce findings or conclusions from a survey not conducted in preparation for litigation or by any of the parties to the litigation. The trial court should take into account the exact issue that the expert seeks to testify about and the nature of the expert's field of

37. The one exception is that sampling expertise would be unnecessary if the survey were administered to all members of the relevant population. *See, e.g.*, McGovern & Lind, *supra* note 27.

38. If survey expertise is being provided by several experts, a single expert may have general familiarity but not special expertise in all these areas.

expertise.[39] All experts who give opinions about the adequacy and interpretation of a survey not only should have general skills and experience with surveys and be familiar with all of the issues addressed in this reference guide, but also should demonstrate familiarity with the following properties of the survey being discussed:

1. Purpose of the survey;
2. Survey methodology,[40] including
   a. the target population,
   b. the sampling design used in conducting the survey,
   c. the survey instrument (questionnaire or interview schedule), and
   d. (for interview surveys) interviewer training and instruction;
3. Results, including rates and patterns of missing data; and
4. Statistical analyses used to interpret the results.

Population Definition and Sampling

The Survey Universe or Population

One of the first steps in designing or evaluating a survey is to identify the target population (or universe).[41] The target population consists of all elements (e.g., individuals) whose characteristics or perceptions the survey is intended to represent. Thus, in trademark litigation, the relevant population in some disputes may include all prospective and past purchasers of the pertinent party's category of

39. For a discussion of the admissibility of expert opinion testimony, see Liesa L. Richter and Daniel J. Capra, *The Admissibility of Expert Testimony*, "Federal Rule of Evidence 702: An Overview" section, in this manual.

40. *See* A & M Records, Inc. v. Napster, Inc., No. C 99–05183 MHP, 2000 U.S. Dist. LEXIS 20668, at *22–25 (N.D. Cal. Aug. 10, 2000) (holding that expert could not attest credibly that the surveys upon which he relied conformed to accepted survey principles because of his minimal role in overseeing the administration of the survey and limited expert report); Hr'g Tr. at 29–30, Munchkin Inc. v. Playtex Prods., LLC, No. CV11-503-AHM(RZx) (C.D. Cal. May 1, 2012), ECF No. 285 (excluding survey when the expert "didn't know how [the survey] was administered[,] . . . didn't know how the panel was selected [and] didn't know what the statistical technique was that was used to weigh the survey and provide weight to it"), *cited in* Cohen v. Trump, No. 3:13-cv-2519-GPC-WVG, 2016 U.S. Dist. LEXIS 117059, at *16–17 (S.D. Cal. Aug. 29, 2016).

41. Identification of the proper target population or universe is recognized uniformly as a key element in the development of a survey. *See, e.g.*, Manual for Complex Litigation, Fourth, § 11.493 (2004), https://www.uscourts.gov/file/3228/download [hereinafter MCL 4th]; *see also* 3 McCarthy, *supra* note 35, § 32:166; Council of Am. Survey Rsch. Orgs., Code of Standards and Ethics for Survey Research § III.A.3 (2011), https://perma.cc/698Z-CF86 [hereinafter CASRO]. (Note that CASRO merged with the Marketing Research Association to form the Insights Association in 2017.)

goods or services. Similarly, the population for a discovery survey may include all potential plaintiffs or all employees who worked for Company A between two specific dates. In a community survey designed to provide evidence for a motion for a change of venue, the relevant population consists of all jury-eligible citizens in the community in which the trial is to take place.[42] The definition of the relevant population is crucial because there may be systematic differences in the responses of members of the population and nonmembers. For example, consumers who are prospective purchasers may know more about the product category than consumers who are not considering making a purchase.

The universe must be defined carefully.[43] For example, in a survey testing whether the defendant made misleading representations about their digital evidence–related software used by police departments, the appropriate universe consisted of potential purchasers of the software in the law enforcement community. Instead, the sample consisted of respondents who merely used photographs in their law enforcement work and had no influence on purchase decisions involving the relevant software or were even necessarily potential users of such software.[44] Defects in the universe may lead to misleading results that should reduce the weight that is given to the survey;[45] a survey should be excluded if it consists of respondents who do not substantially reflect the characteristics of the relevant population.[46]

42. An additional relevant population may consist of jury-eligible citizens in the community where the party would like to see the trial moved. By questioning citizens in more than one community, the survey can test whether moving the trial is likely to reduce the level of animosity toward the party requesting the change of venue. *See* United States v. Cloud, No. 1:19-cr-02032-SMJ-1, 2021 U.S. Dist. LEXIS 260839, at *8–9 (E.D. Wash. Dec. 8, 2021) (order granting a motion for intra-district transfer from Yakima based in part on a telephone survey of jury-eligible respondents in Yakima, Richland, and Spokane); United States v. Haldeman, 559 F.2d 31, 140, 151 (MacKinnon, J., dissenting), 176–79 (app. A) (D.C. Cir. 1976) (court denied change of venue over the strong objection of Judge MacKinnon, who cited survey evidence that Washington, D.C., residents were substantially more likely to conclude, before trial, that the defendants were guilty); *see also* People v. Venegas, 31 Cal. Rptr. 2d 114, 117 (Cal. Ct. App. 1994) (change of venue denied because defendant failed to show that the defendant would face a less hostile jury in a different court).

43. *See* Merck Eprova AG v. Brookstone Pharms., 920 F. Supp. 2d 404, 418 (S.D.N.Y. 2013) ("In determining whether a challenged advertisement is likely to confuse or mislead customers, courts must look to the person to whom the advertisement is addressed.") (citation omitted).

44. *See, e.g.*, Kwan Software Eng'g, Inc. v. Foray Techs., LLC, 110 U.S.P.Q.2d 1637, 1641–42 (N.D. Cal. 2014). *See also* Warner Bros., Inc. v. Gay Toys, Inc., 658 F.2d 76 (2d Cir. 1981) (surveying child users of the product rather than parent purchasers). Children and some other populations create special challenges for researchers. For example, very young children should not be asked about sponsorship or licensing, concepts that are foreign to them. Concepts, as well as wording, should be age appropriate.

45. *See, e.g.*, Chi. Mercantile Exchange Inc. v. ICE Clear US, Inc., No. 18 C 1376, 2020 WL 1905760, at *12–14 (N.D. Ill. Apr. 12, 2020).

46. *See, e.g.*, *Kwan*, 110 U.S.P.Q.2d at 1642.

The Sampling Frame as an Approximation of the Population

The target population consists of all the individuals or units whose responses the researcher would like to describe. The sampling frame is the source (or sources) from which the sample actually is drawn. The surveyor's job generally is easier if a complete list of every eligible member of the population is available (e.g., all plaintiffs in a discovery survey), so that the sampling frame lists all members of the target population. The survey expert should identify how the sampling frame was compiled—that is, the source(s) used to obtain the list of members of the population. Ideally, even if a list of every member of the population is not available, the survey expert will still have access to demographic characteristics of the full population (e.g., sex, race/ethnicity, age). In some cases, this is straightforward, such as when the population consists of all American residents (so the Census's American Community Survey can provide the demographic information). In other situations, population descriptions may be less easily obtained (e.g., the consumers of a particular product, whose demographic characteristics may be identifiable only from the marketing research of the company that produces the product). In practice, the target population often includes some members who cannot be contacted or who cannot be identified in advance. As a result, reasonable compromises are sometimes required in developing the sampling frame. The survey report should contain (1) a description of the target population, (2) a description of the sampling frame from which the sample is drawn, (3) a discussion of the difference between the two, and, importantly, (4) an evaluation of the likely consequences of that difference.

A survey that provides information about a wholly irrelevant population is itself irrelevant.[47] Courts are likely to exclude such a survey or accord it minimal

47. A survey aimed at assessing how persons in the trade respond to an advertisement should be conducted on a sample of persons in the trade and not on a sample of consumers. *See* Home Box Off. v. Showtime/The Movie Channel, 665 F. Supp. 1079, 1083–84 (S.D.N.Y. 1987), *aff'd in part and vacated in part*, 832 F.2d 1311 (2d Cir. 1987); J & J Snack Food Corp. v. Earthgrains Co., 220 F. Supp. 2d 358, 371–72 (D.N.J. 2002); Parks, LLC v. Tyson Foods, Inc., 186 F. Supp. 3d 405, 419–20 (E.D. Pa. 2016) (giving little weight to a survey aimed at consumers of the plaintiff's products when it should have been targeted at users of the defendant's). *But see* Lon Tai Shing Co., LTD v. Koch & Lowy, No. 90 Civ. 4464, 1990 U.S. Dist. LEXIS 19123, at *50–57 (S.D.N.Y. Dec. 14, 1990), in which the judge was willing to find likelihood of consumer confusion from a survey of lighting store salespersons questioned by a survey researcher posing as a customer. The court was persuaded that the salespersons who were misstating the source of the lamp, whether consciously or not, must have reasonably believed that the consuming public would be likely to rely on the salespersons' inaccurate statements about the name of the company that manufactured the lamp they were selling.

weight.[48] More commonly, however, the sampling frame and the target population have some overlap, but the overlap is imperfect. This is called coverage error, and it has two types: the sampling frame may be underinclusive by excluding part of the target population, or it may be overinclusive by including individuals who are not members of the target population. If the coverage is underinclusive, the survey's value depends on the extent to which the excluded population is likely to respond differently from the included population. Thus, a survey of spectators and participants at running events would be sampling a sophisticated subset of those likely to purchase running shoes. Because this subset would probably consist of the consumers most knowledgeable about the trade dress used by companies that sell running shoes, a survey based on this sampling frame would be likely to substantially overrepresent the strength of a particular design as a trademark. Worse still, the extent of that overrepresentation would be unknown and not susceptible to any reasonable estimation.[49]

In some cases, it is difficult to determine whether a sampling frame that omits some members of the population distorts the results of the survey and, if so, the extent and likely direction of the bias. For example, a trademark survey was designed to test the likelihood of confusing an analgesic currently on the market with a new product that was similar in appearance.[50] The plaintiff's survey included only respondents who had used the plaintiff's analgesic, and the court found that the target population should have included users of other analgesics, "so that the full range of potential customers for whom plaintiff and defendants would compete could be studied."[51] In this instance, it is unclear whether users of the plaintiff's product would be more or less likely to be confused than users of the defendants' product or users of a third analgesic, but the omission of users

48. *See* Varner v. Dometic Corp., No. 16-22482-CIV-SCOLA/OTAZO-REYES, 2022 U.S. Dist. LEXIS 127353, at *27–30 (S.D. Fla. Feb. 2, 2022) (survey of consumers excluded because gas-absorption refrigerators are not sold directly to consumers, but rather to manufacturers and OEM and RV dealers); *see also In re* Fluidmaster, Inc., Water Connector Components Prods. Liab. Litig., No. 14-cv-5696, 2017 WL 1196990, at *29 (N.D. Ill. Mar. 21, 2017); Wells Fargo & Co. v. WhenU .com, Inc., 293 F. Supp. 2d 734 (E.D. Mich. 2003).

49. *See* Brooks Shoe Mfg. Co. v. Suave Shoe Corp., 533 F. Supp. 75, 80 (S.D. Fla. 1981), *aff'd*, 716 F.2d 854 (11th Cir. 1983); *see also* Hodgdon Powder Co. v. Alliant Techsystems, Inc., 512 F. Supp. 2d 1178, 1181–82 (D. Kan. 2007) (excluding survey on gunpowder brands distributed at plaintiff's promotional booth at a shooting tournament); Winning Ways, Inc. v. Holloway Sportswear, Inc., 913 F. Supp. 1454, 1467 (D. Kan. 1996) (survey flawed in failing to include sporting goods customers who constituted a major portion of customers). *But see* Thomas & Betts Corp. v. Panduit Corp., 138 F.3d 277, 294–95 (7th Cir. 1998) (survey of store personnel admissible because relevant market included both distributors and ultimate purchasers).

50. *See* Am. Home Prods. Corp. v. Barr Lab'ys, Inc., 656 F. Supp. 1058 (D.N.J.), *aff'd*, 834 F.2d 368 (3d Cir. 1987).

51. *Am. Home Prods.*, 656 F. Supp. at 1070.

of other analgesics made it impossible to assess the effect of that difference.[52] If the sampling frame does not include important groups in the target population, the survey cannot provide information on how the unrepresented members of the target population would have responded.[53]

An overinclusive sampling frame generally presents less of a problem for interpretation than does an underinclusive sampling frame.[54] If the survey expert can demonstrate that a sufficiently large and representative subset of respondents in the survey was drawn from the appropriate sampling frame, the responses obtained from that subset can be examined, and inferences about the relevant population can be drawn based on that subset.[55] If the relevant subset cannot be identified, however, an overbroad sampling frame will reduce (or eliminate) the value of the survey.[56]

The Sample as a Reflection of the Relevant Characteristics of the Population

Identification of a survey population must be followed by selection of a sample that accurately represents that population.[57] The use of probability sampling techniques maximizes both the representativeness of the survey results and the

52. *See also* Craig v. Boren, 429 U.S. 190 (1976).

53. *See, e.g.*, Amstar Corp. v. Domino's Pizza, Inc., 615 F.2d 252, 263–64 (5th Cir. 1980) (court found both plaintiff's and defendant's surveys substantially defective for a systematic failure to include parts of the relevant population); Scott Fetzer Co. v. House of Vacuums, Inc., 381 F.3d 477, 487–88 (5th Cir. 2004) (universe drawn from plaintiff's customer list is underinclusive and customers are likely to differ in their familiarity with plaintiff's marketing and distribution techniques); Hi Ltd. P'Ship v. Winghouse of Fla., Inc., No. 6:03-cv-116-Orl-22JGG, 2004 U.S. Dist. LEXIS 30687, at *25–26 (M.D. Fla. Oct. 5, 2004) ("[T]he failure to include a single female in the survey, when women comprise nearly a third of Hooters' customer base and perhaps even more of Winghouse's clientele, reflects a patently flawed methodology striking at the very heart of the survey's validity.").

54. *See* Schwab v. Philip Morris USA, Inc., 449 F. Supp. 2d 992, 1135 (E.D.N.Y. 2006) ("Studies evaluating broadly the beliefs of low tar smokers generally are relevant to the beliefs of 'light' smokers more specifically."); Jacobs v. Fareportal, Inc., No. 8:17CV362, 2020 U.S. Dist. LEXIS 211840, at *25–26 (D. Neb. May 29, 2020) (critique about the overinclusiveness of sampling past as well as future purchasers goes to weight rather than admissibility).

55. *See* Nat'l Football League Props., Inc. v. Wichita Falls Sportswear, Inc., 532 F. Supp. 651, 657–58 (W.D. Wash. 1982).

56. *See* Leelanau Wine Cellars, Ltd. v. Black & Red, Inc., 502 F.3d 504, 518 (6th Cir. 2007) (lower court was correct in giving little weight to survey with overbroad universe); Big Dog Motorcycles, L.L.C. v. Big Dog Holdings, Inc., 402 F. Supp. 2d 1312, 1334 (D. Kan. 2005) (universe composed of prospective purchasers of all t-shirts and caps overinclusive for evaluating reactions of buyers likely to purchase merchandise at motorcycle dealerships). *See also* Schieffelin & Co. v. Jack Co. of Boca, Inc., 850 F. Supp. 232, 246 (S.D.N.Y. 1994).

57. MCL 4th, *supra* note 41, § 11.493.

ability to assess the precision of estimates obtained from the survey. Probability samples range from simple random samples to complex multistage sampling designs that use stratification, clustering of population elements into various groupings, or both. In all forms of probability sampling, each element in the relevant population has a known, nonzero probability of being included in the sample.[58] These sample selection probabilities do not need to be the same for all population elements (i.e., some members may have a higher or lower chance of being selected than others). If the probabilities are unequal, however, compensatory adjustments should be made in the analysis. Failure to adjust by weighting to reflect the known distribution in the population on relevant characteristics may undermine representativeness and warrant exclusion of the survey.[59]

Probability sampling offers two important advantages over nonprobability sampling. First, a probability sample can provide an unbiased estimate that summarizes the responses of all persons in the population from which the sample was drawn; that is, the expected value of the sample estimate is the population value being estimated. For instance, if a probability sample leads to an estimate that 80% of respondents were confused as to whether an analgesic was an existing or new option, the researcher could be confident that approximately 80% of those in the population would have that belief (within some margin of error). A probability sample can provide an unbiased estimate of the population even when the size of the sample is relatively small. The advantage of larger samples is that they provide more precision (smaller margins of error). In general, the absolute size of the sample, rather than its size relative to the population, determines the precision of the estimate.[60]

Second, probability sampling allows the researcher to calculate a confidence interval that explicitly provides information on the reliability of the sample estimate of the population value (i.e., the value one would obtain if everyone in the population responded). The difference between the estimate and the exact value

58. *See* Thomas Piazza, *Fundamentals of Applied Sampling*, *in* Handbook of Survey Research, *supra* note 1, at 139, 145.

59. Fish v. Kobach, 309 F. Supp. 3d 1048, 1059–61 (D. Kan. 2018), *aff'd sub nom.* Fish v. Schwab, No. 18–3133, 2020 U.S. App. LEXIS 13723 (10th Cir. Apr. 29, 2020) (expert's survey intended to represent the eligible voting population of Kansas was excluded, in part, for failure to weight responses to reflect the distribution of educational attainment and household income level in the population).

60. An exception to this rule applies when a population is (identifiably) finite and the sample size is large relative to the population size, typically taken to occur when the sample size is greater than 5% of the total population (e.g., the population strictly includes 1,000 people, and the sample includes more than 50 people). In this case, it is appropriate to adjust the standard error/margin of error by the finite population correction (FPC) factor. The FPC factor incorporates both the size of the sample and its share of the population in determining the degree of precision. *See* William G. Cochran, Sampling Techniques 24–25 (3d ed. 1977).

is called the sampling error.[61] Thus, suppose a survey collected responses from a simple random sample of 400 dentists selected from the population of all dentists licensed to practice in the United States and found that 20% of them mistakenly believed that a new toothpaste, Goldgate, was manufactured by the makers of Colgate. A survey expert could properly compute a confidence interval around the 20% estimate obtained from this sample. If the survey were repeated a large number of times, and a 95% confidence interval was computed each time, 95% of the confidence intervals would include the actual percentage of dentists in the entire population who would believe that Goldgate was manufactured by the makers of Colgate.[62] In this example, the margin of error is ±4%, and so the confidence interval is the range between 16% and 24%—that is, the estimate (20%) plus or minus 4%.

All sample surveys (i.e., surveys that do not measure responses from every member of the population) produce estimates of population values, not exact measures of those values. Assuming a probability sample, a confidence interval describes how stable the mean response in the sample is likely to be. The width of the confidence interval depends on three primary characteristics:

1. Size of the sample (larger samples produce narrower intervals);
2. Variability of the response being measured (more variability leads to wider intervals); and
3. Confidence level the researcher wants to have.[63]

Traditionally, scientists adopt the 95% level of confidence, which means that if one hundred samples of the same size were drawn, the confidence interval expected for ninety-five of the samples would be expected to include the true population value.[64]

Stratified probability sampling uses what is known about the population characteristics to correct for imbalances produced by random sampling. Consider the dentist example presented earlier: if it is known that 60% of dentists have at least

61. *See* the glossary of this chapter, and David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual, for a more detailed definition of sampling error.

62. Actually, because survey interviewers would be unable to locate some dentists, and some dentists would be unwilling to participate in the survey, technically the population to which this sample would be projectable would be all dentists with current addresses who would be willing to participate in the survey if they were asked. The expert should be prepared to discuss possible sources of bias due to, for example, an address list that is not current.

63. When the sample design does not use a simple random sample, the confidence interval will be affected.

64. To increase the likelihood that the confidence interval contains the actual population value (e.g., from 95% to 99%) without increasing the sample size, the width of the confidence interval can be expanded. An increase in the confidence interval brings an increase in the confidence level. For further discussion of confidence intervals, see the glossary to the current chapter, and David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods,* in this manual.

fifteen years of experience and 40% have fewer years, the researcher could randomly sample dentists within each of those separate categories to obtain a sample that reflected those proportions. This would ensure the sample was unbiased with respect to years of experience—that proportion would be set by the researcher—and would therefore reduce sampling error.[65] Disproportionate sampling from subgroups may be used to enable the survey to provide separate estimates for particular subgroups but should be weighted (as discussed below) to reflect the population as a whole when conducting an overall analysis.

The last decade has seen a dramatic rise in the availability of nonprobability samples. Most of these samples come in one of three general varieties. The first variety is drawn from large nonprobability internet panel samples overseen by a vendor. For these panels, individuals opt in (e.g., via advertisements) to receive compensation for taking surveys. The companies that oversee these panels can produce a set of respondents that consists of members of the target population (e.g., the population of elementary school teachers) and matches the distribution of that population on specified characteristics (e.g., percentage of women and minorities). These samples vary in their ability to hit benchmarks, but some perform quite well.[66] Regardless, these samples differ from probability samples because they are not drawn from a list of the entire population. Instead, there is an attempt to identify relevant respondents and to balance the sample to resemble the population on certain observable features (e.g., the sample and population have the same distribution of gender, age, income).

The second source of nonprobability samples is crowdsourcing labor market platforms, the best-known of which is Amazon's Mechanical Turk (MTurk). Researchers can directly use MTurk or analogous platforms to hire individuals to complete tasks, including taking surveys, for direct compensation. Here, researchers can try to draw samples that match populations on observable variables, but it is often difficult since these platforms do not easily allow for the kind of selective participant invitations that permit managed panels to build custom samples.

The third source of nonprobability samples are purposive samples. These use nonprobability methods designed to target hard-to-reach populations (for whom probability sampling is extremely difficult or impossible), such as low-income people, young people, Indigenous people, and people in poor health.[67] These sampling methods are sometimes the subject of judicial concerns about researcher bias in selecting respondents, as these methods require the researcher to engage

65. *See* Pharmacia Corp. v. Alcon Lab'ys, Inc., 201 F. Supp. 2d 335, 365 (D.N.J. 2002).

66. Lynn Vavreck & Douglas Rivers, *The 2006 Cooperative Congressional Election Study*, 18 J. Elections, Pub. Op., & Parties, 355–66 (2008), https://doi.org/10.1080/17457280802305177.

67. Hard-to-Survey Populations (Roger Tourangeau et al. eds., 2014). Abdolreza Shaghaghi et al., *Approaches to Recruiting 'Hard-To-Reach' Populations into Research: A Review of the Literature*, 1 Health Promotion Persps. 86, 94 (2011), https://doi.org/10.5681/hpp.2011.009.

in active and targeted outreach.[68] They also require particular attention to language and survey mode.[69]

Generally speaking, researchers employ nonprobability samples because they are substantially less expensive, often allow for larger samples (due to the lower cost), may include a higher number of individuals from less prevalent subgroups (e.g., members of racial or ethnic minority groups) that could be of interest, and/or offer the only option for hard-to-reach populations. Indeed, the use of nonprobability samples has been common in litigation for some time. For instance, an overwhelming majority of the consumer surveys conducted for Lanham Act litigation present results from nonprobability samples,[70] and the standard modern practice is for these surveys to be conducted online using professionally managed nonprobability internet panels (i.e., the first type discussed above).[71] A common rationale for admitting such surveys into evidence is that they are used widely in marketing research and that "results of these studies are used by major American companies in making decisions of considerable consequence."[72]

Nonprobability samples are appropriate for some types of litigation surveys but still carry important limitations. In many cases, researchers cannot compute response rates for nonprobability samples because they are unaware of how many potential respondents viewed the invitation to participate in the survey and decided not to do so (e.g., researchers often just post participation invitations to those in a labor market without knowing how many people viewed it). In contrast, when trying to assess the base rate of some variable in the national population, probability samples offer superior accuracy relative to nonprobability samples

68. *See, e.g.*, Branson v. All. Coal, LLC, No. 4:19-CV-00155-JHM-HBB, 2022 U.S. Dist. LEXIS 123731, at *23–24 (W.D. Ky. July 13, 2022) (rejecting the use of purposive sampling in a class certification survey: "Put simply, the inclusion of purposefully selected data questions whether the data accurately represents the population as a whole. To preserve the reliability of the discovery data, no party shall select certain opt-in plaintiffs from whom to collect information or depose.").

69. Robin Bayes et al., *Studying Science Inequities: How to Use Surveys to Study Diverse Populations,* 700 Annals Am. Acad. Pol. & Soc. Sci., 220, 233 (2022), https://doi.org/10.1177/00027162221093970.

70. Jacob Jacoby & Amy H. Handlin, *Non-Probability Sampling Designs for Litigation Surveys*, 81 Trademark Rep. 169, 173 (1991). For probability surveys conducted in trademark cases, see *James Burrough, Ltd. v. Sign of Beefeater, Inc.*, 540 F.2d 266 (7th Cir. 1976); *Nightlight Systems v. Nitelites Franchise Systems*, No. 1:04-CV-2112-CAP, 2007 U.S. Dist. LEXIS 95565 (N.D. Ga. July 17, 2007); *National Football League Properties, Inc. v. Wichita Falls Sportswear, Inc.*, 532 F. Supp. 651 (W.D. Wash. 1982).

71. Matthew B. Kugler & R. Charles Henn, *Internet Surveys in Trademark Cases: Benefits, Challenges, and Solutions*, *in* Trademark and Deceptive Advertising Surveys 291, 293 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

72. Nat'l Football League Props., Inc. v. N.J. Giants, Inc., 637 F. Supp. 507, 515 (D.N.J. 1986). A survey of members of the Council of American Survey Research Organizations, the national trade association for commercial survey research firms in the United States, revealed that 95% of the in-person independent contacts in studies done in 1985 took place in malls or shopping centers. Jacoby & Handlin, *supra* note 70, at 172–73, 176.

on validated demographic measures such as sex, age, race, and ethnicity,[73] as well as validated behavioral measures such as vaccine uptake.[74] This is because probability samples are more representative of the underlying population. In evaluating the representativeness of a nonprobability sample relative to the population, the expert should consider both observable and nonobservable sample demographics. On critical variables, it may be necessary to include the joint distributions of the measured variables (e.g., the percentage of minority women).[75] More generally, there is always the possibility of unobserved relevant factors even in a well-conducted nonprobability survey.[76] For instance, a nonprobability sample with appropriate quotas may still not include typical members from each quota group. This can lead to inaccuracies in point estimates. Nonprobability samples can provide unbiased estimates if the sample is representative on all variables that correlate with the outcome of interest, or has been weighted to be representative, but this can be challenging when it comes to unobserved factors.[77]

Ultimately, the importance of using a probability sample depends on the survey's goal. If the survey is designed to make a causal inference based on differences between randomly assigned experimental conditions (40% of people in

73. Bo MacInnis et al., *The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension*, 82 Pub. Op. Q. 707 (2018), https://doi.org/10.1093/poq/nfy038. The authors compare the accuracy of probability surveys (RDD telephone and internet), internet surveys that combine nonprobability and probability samples, and internet surveys that are fully nonprobability (but with different types of compensation). This study is the most extensive of its kind, using a set of fifty measures and forty benchmark variables from federal face-to-face surveys with high response rates. The analyses showed that probability samples provide more accurate estimates of population quantities than nonprobability samples as well as samples that combined methods.

74. Valerie C. Bradley et al., *Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake*, 600 Nature 695 (2021), https://doi.org/10.1038/s41586-021-04198-4.

75. Connor Huff & Dustin Tingley, *"Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents*, 2 Rsch. & Politics 1 (2015), https://doi.org/10.1177/2053168015604648.

76. The court in *Kinetic Concepts, Inc. v. BlueSky Medical Corp.*, No. SA-03-CA-0832, 2006 U.S. Dist. LEXIS 60187, at *14–17 (W.D. Tex. Aug. 11, 2006), found the plaintiff's survey using a nonprobability sample to be admissible and permitted the plaintiff's expert to present results from a survey using a convenience sample. The court then assisted the jury by providing an instruction on the differences between probability and convenience samples and the estimates obtained from each.

77. *See, e.g.*, Robert M. Groves et al., Survey Methodology (2d ed. 2009). Groves and colleagues point out that using statistical significance tests and confidence intervals with nonprobability samples is technically inappropriate since it becomes impossible to obtain an unbiased estimate of sampling variance. That said, they accept that nonprobability samples can approximate probability samples when the researcher knows what population attributes correlate with the key statistics and ensures that the sample is balanced on those attributes (i.e., using quotas). *Id.* at 409–10. Vasia Vehovar and colleagues agree: "We thus recommend to be more openly accepting of the reality of using a standard statistical inference approach as an approximation in non-probability settings" as long as the nature of how the sample is drawn is clear. Vasia Vehovar et al., *Non-Probability Sampling*, *in* The SAGE Handbook of Survey Methodology (2016), https://doi.org/10.4135/9781473957893.

the experimental condition were confused by defendant's advertising, but only 20% by control advertising), a survey-experiment using a nonprobability sample of relevant respondents will generally suffice. If the aim is to obtain a precise measure of a particular feature of a population, however, a nonprobability sample faces challenges.[78] An expert should consider the potential biases in the sample and whether they impact factors correlated with the variable being estimated, to assess the likely magnitude of the biases (i.e., how much they are likely to affect the estimates). The expert should be prepared to explain how observable and unobservable sample characteristics might impact their estimate.

When using a nonprobability sample in an experimental context (i.e., a survey-experiment), the primary concern about the characteristics of the sample is whether they modify the causal relationship being tested in the survey.[79] For instance, if more and less experienced dentists react the same way to the name Goldgate for toothpaste—if experience does not moderate (that is, change) the effect of the name—a survey-experiment that does not include more experienced dentists will still produce results that generalize across experience level. In contrast, if reaction to the name does vary with experience—if experience *does* moderate the reaction—a sample that includes only less experienced dentists will produce a biased result. The survey expert should identify any potential sample characteristics that may moderate the effects being assessed and demonstrate that the sampling approach takes them into account.

In sum, the historic gold standard was a probability sample because it could provide unbiased estimates of the population. Yet, increased costs of probability samples, more ways to obtain nonprobability samples, challenges with hard-to-reach populations, and nonresponse bias (discussed below) have all led to greater use of nonprobability samples. When such samples are used, it is vital to clarify how the sample compares to known-probability benchmarks (e.g., demographic

78. However, even if the initial sampling approach uses a probability sample, systematic nonresponse can undermine the representativeness of the sample and hence the accuracy of a precise estimate. This effect has been demonstrated by studies of the effect of a two-stage process to recruit survey participants in which the initial invitation used probability sampling through U.S. Postal Service mailings, phone contact, and modest incentives, and a follow-up effort involving FedEx mailings, in-person recruitment by field interviewers, and enhanced incentives attempted to recruit initial nonresponders. Comparisons of the two groups indicated small but statistically significant differences in reported political views and income level. The importance of such differences will depend on the claim being made for the accuracy of the estimate. Ipek Bilgen et al., The Undercounted: Measuring the Impact of 'Nonresponse Follow-up' on Research Data (2019), https://perma.cc/P5RS-268Q.

79. *See* James N. Druckman & Cindy D. Kam, *Students as Experimental Participants: A Defense of the 'Narrow Data Base'*, *in* Cambridge Handbook of Experimental Political Science (2011); James N. Druckman, Experimental Thinking: A Primer on Social Science Experiments (2022).

characteristics and other features that could influence the outcome being studied), what adjustments may have been made (e.g., weights; see below) and how biases may influence inferences—particularly when the goal is to arrive at a precise estimate of a population value. Nonprobability samples tend to be on stronger footing when used in an experimental context, but even here it is vital for the researcher to discuss potential variables that could moderate the experimental effect and whether the sampling approach influenced the impact of those moderators.

Nonresponse as a Potential Source of Bias

Even when a sample is drawn randomly from a complete list of elements in the target population (i.e., a probability sample), responses or measures are typically obtained from only part of the sample. If this lack of response is distributed randomly, valid inferences about the population can be drawn with assurance using the measures obtained from the available sample. But nonresponse often is not random. For example, persons who are single typically have three times the "not at home" rate in U.S. Census Bureau surveys as do people living with family.[80] Nonresponse also occurs when contact with the sampled individual is made, but that person declines to participate. This problem arose recently in political polls where Republicans seemed less likely to participate than Democrats.[81] Efforts to increase response rates include making several attempts to contact potential respondents, sending advance letters,[82] and providing financial or nonmonetary incentives for participating in the survey.[83]

The key to evaluating the effect of nonresponse in a survey is to determine the extent to which nonrespondents differ from respondents in a way that would impact the results of the survey. If nonresponse has biased the pattern of responses, the size and direction of that bias need to be assessed. On some occasions, it may be possible to anticipate systematic patterns of nonresponse. For example, high-volume medical professionals may be less willing to respond to a survey than those with lower-volume practices. If the survey includes questions about volume of

80. 2 Gastwirth, *supra* note 30, at 501.

81. Courtney Kennedy et al., Pew Research Center, Confronting 2016 and 2020 Polling Limitations (2021), https://perma.cc/P5CC-HJH2; *see also* Joshua D. Clinton et al., *Reluctant Republicans, Eager Democrats? Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-election Telephone Polls*, 86 Pub. Op. Q. 247 (2022), https://doi.org/10.1093/poq/nfac011.

82. Edith De Leeuw et al., *The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis*, 71 Pub. Op. Q. 413 (2007), https://doi.org/10.1093/poq/nfm014 (advance letters effective in increasing response rates in telephone as well as mail and face-to-face surveys).

83. Erica Ryu et al., *Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error*, 18 Int'l J. Pub. Op. Rsch. 89 (2006), https://doi.org/10.1093/ijpor/edh089.

practice, the expert can assess how experience level may have affected the pattern of results.[84]

Although high response rates are desirable because they reduce the potential impact of nonresponse bias, they are increasingly difficult to achieve. Survey nonresponse rates have risen substantially in the twenty-first century, along with the costs of obtaining responses, and so the issue of nonresponse has attracted considerable attention from survey researchers.[85]

Researchers have developed a variety of approaches to adjust for nonresponse, including weighting obtained responses in proportion to known demographic characteristics of the target population (which we discuss below), comparing the pattern of responses from early and late responders to mail surveys or the pattern of responses from easy-to-reach and hard-to-reach responders in telephone surveys, and imputing estimated responses to nonrespondents based on known characteristics of those who have responded. All of these techniques can only approximate the response patterns that would have been obtained if nonrespondents had responded. Nonetheless, they are useful for testing the robustness of the estimates obtained from responders.

To assess the general impact of the lower response rates, researchers have compared results obtained from surveys with varying response rates.[86] Surprisingly comparable results have been obtained in many surveys with varying response rates, suggesting that surveys may achieve reasonable estimates even with relatively low response rates. The key is whether nonresponse is associated with systematic differences in response that cannot be adequately modeled or assessed. Generally, the representativeness of a sample (regardless of response rate) matters much more for accurate inference than the response rate.[87] Determining whether

84. In *People v. Williams*, 830 N.Y.S.2d 452 (N.Y. Sup. Ct. 2006), a published survey of experts in eyewitness research was used to show general acceptance of various eyewitness phenomena. *See* Saul Kassin et al., *On the "General Acceptance" of Eyewitness Testimony Research: A New Survey of the Experts*, 56 Am. Psychologist 405 (2001), https://doi.org/10.1037/0003-066X.56.5.405. The survey included questions on the publication activity of respondents and compared the responses of those with high and low research productivity. Productivity levels in the respondent sample suggested that respondents constituted a blue-ribbon group of leading researchers. *Williams*, 830 N.Y.S.2d at 457, 459 n.26. *See also* Pharmacia Corp. v. Alcon Lab'ys, Inc., 201 F. Supp. 2d 335 (D.N.J. 2002).

85. *E.g.*, Richard Curtin et al., *Changes in Telephone Survey Nonresponse over the Past Quarter Century*, 69 Pub. Op. Q. 87 (2005), https://doi.org/10.1093/poq/nfi002. *See also* Robert M. Groves & Emilia Peytcheva, *The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis*, 72 Pub. Op. Q. 167 (2008), https://doi.org/10.1093/poq/nfn011; Peter Miller et al., Federal Committee on Statistical Methodology, A Systematic Review of Nonresponse Bias Studies in Federally Sponsored Surveys (2020), https://perma.cc/6WCQ-AFNK.

86. *E.g.*, Daniel M. Merkle & Murray Edelman, *Nonresponse in Exit Polls: A Comprehensive Analysis*, *in* Survey Nonresponse 243–57 (Robert M. Groves et al. eds., 2002) (finding minimal nonresponse error associated with refusals to participate in in-person exit polls); *see also* Jon A. Krosnick, *Survey Research*, 50 Ann. Rev. Psych. 537 (1999), https://doi.org/10.1146/annurev.psych.50.1.537.

87. Bo MacInnis et al., *The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension*, 82 Pub. Op. Q. 707 (2018), https://doi.org/10.1093/poq/nfy038;

the level of nonresponse in a survey seriously impairs inferences drawn from the results generally requires an analysis of the determinants of nonresponse. For example, even a survey with a high response rate may seriously underrepresent some portions of the population, such as the unemployed or the poor. The survey expert should be prepared to provide evidence on the potential impact of nonresponse on the survey results.

In surveys that include sensitive or difficult questions, some respondents may refuse to provide answers or may provide incomplete answers.[88] To assess the impact of nonresponse to a particular question, the survey expert should analyze the differences between those who answered and those who did not answer. Procedures to address the problem of missing data include recontacting respondents to obtain the missing answers and using a respondent's other answers to predict the missing response (i.e., imputation).[89]

Survey researchers commonly use weights, even with many probability samples, to address coverage gaps and control for differential nonresponse among subgroups. Weighting adjustments can sometimes improve the accuracy of survey results.[90] Responses are weighted to reflect distributions in the target population, typically demographics.[91] When the population includes all Americans, the U.S. Census or American Community Survey can provide demographic population figures that can guide how to weight the survey responses. For instance, if the population consists of 50% men but the sample contains only 40% men, then male sample respondents will be weighted to count more in computations from the sample (and women will be counted less) to better approximate the target

Scott Keeter, *Evidence About the Accuracy of Surveys in the Face of Declining Response Rates*, *in* The Palgrave Handbook of Survey Research 19–22 (David L. Vannette & Jon A. Krosnick eds., 2018), https://doi.org/10.1007/978-3-319-54395-6_4.

88. *See* Roger Tourangeau at al., The Psychology of Survey Response (2000); California v. Ross, 358 F. Supp. 3d 965 (N.D. Cal. 2019) (surveys were used to show that if a question on citizenship was included in the 2020 census, some individuals would be less likely to participate or answer the citizenship question, and that this refusal was likely to be higher in the Latinx community).

89. *See* Paul D. Allison, *Missing Data*, *in* Handbook of Survey Research, *supra* note 1, at 630; *see also* Groves & Peytcheva, *supra* note 85.

90. Heidi Jensen et al., *The Impact of Non-Response Weighting in Health Surveys for Estimates on Primary Health Care Utilization*, 32 Eur. J. Public Health 450 (2022), https://doi.org/10.1093/eurpub/ckac032. *But see* Kenneth Bollen et al., *Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis*, 3 Ann. Rev. Stat. Application 375 (2016), https://doi.org/10.1146/annurev-statistics-011516-012958 (suggesting that weighting is sometimes unhelpful and inefficient).

91. Jelke Bethlehem & Mario Callegaro, *Part IV Weighting Adjustments*, *in* Online Panel Research: A Data Quality Perspective 264–72 (Mario Callegaro et al. eds., 2014). *See also* Fish v. Kobach, 309 F. Supp. 3d 1048, 1089 (D. Kan. 2018) (criticizing an expert for not assigning weights to their results given the differences between their sample's demographics and the expected population demographics); California v. Ross, 358 F. Supp. 3d 965, 984 (N.D. Cal. 2019) (commenting that the expert used weighting to balance the demographics of the sample).

population.[92] When responses are weighted, it is crucial to be transparent about how the data were weighted, and to acknowledge that weighting reduces statistical power and, hence, the precision of the estimates.[93]

There are three main challenges with survey weights. First, the survey expert needs to know the true proportions in the population (how many are men, aged 18–35, college educated, etc.), or the sample weights will be largely guesses. Weighting cannot compensate for lack of surveyor knowledge about the underlying distributions in the population. Second, weighting does not actually increase sample size. If a sample of forty African American respondents is weighted at 1.5 (meaning each counts as one and a half people for purposes of analysis), the size of their subsample (and consequently the number used to calculate the margin of error of that subsample) is still forty, not sixty. In fact, weighting decreases the precision of estimates, leading to larger confidence intervals. Finally, weighting cannot correct for unobserved factors. If the survey has recruited unrepresentative members of a subpopulation (for example, atypical Republicans), weighting cannot somehow make them representative. This may be a particular problem when weighting is used to substantially amplify the responses of a small number of people.[94]

Precautions Taken to Ensure That Only Qualified Respondents Were Included in the Survey

In a carefully executed survey, each potential respondent is questioned on the attributes that determine their eligibility to participate in the survey. Thus, the initial questions screen potential respondents to determine if they are members of the target population of the survey (e.g., Does she own a dog? Does he live within ten miles of where he works?). The screening questions must be drafted so that they do not appeal to or deter specific groups within the target population or convey information that will influence a respondent's answers on the main survey (i.e., create a context effect). For example, if respondents must be prospective or recent purchasers of Sunshine orange juice in a trademark survey

92. The process often involves iteratively adjusting for multiple demographics. The idea is to assign an adjustment weight to each respondent such that those from underrepresented groups receive weights greater than 1 and those from overrepresented groups receive weights less than 1. There are a host of ways to weight, and caps should be placed on how much a given observation can be weighted so that one respondent does not have a grossly disproportionate effect on the outcomes. *See* Matthew DeBell, *Computation of Survey Weights*, *in* The Palgrave Handbook of Survey Research 519–27 (2018).

93. Thomas Piazza, *Fundamentals of Applied Sampling*, in Handbook of Survey Research, *supra* note 1. *See also* Texas v. Holder, 888 F. Supp. 2d 113, 131 (D.D.C. 2012) (criticizing an expert for inconsistent use of weighting).

94. Nate Cohn, *How One 19-Year-Old Illinois Man Is Distorting National Polling Averages*, N.Y. Time, Oct. 12, 2016, https://perma.cc/76R7-J363.

designed to assess consumer confusion with Sun Time orange juice, potential respondents might be asked to name the brands of orange juice they have purchased recently or expect to purchase in the next six months. They should not be asked specifically if they recently have purchased, or expect to purchase, Sunshine orange juice, because this may affect their responses on the survey either by implying who is conducting the survey or by supplying them with a brand name that otherwise would not occur to them.

The content of a screening questionnaire (or "screener") can also set the context for the questions that follow. In *Pfizer, Inc. v. Astra Pharmaceutical Products, Inc.*,[95] physicians were asked a screening question to determine whether they prescribed particular drugs. The survey question that followed the screener asked, "Thinking of the practice of cardiovascular medicine, what first comes to mind when you hear the letters XL?" The court found that the screener conditioned the physicians to respond with the name of a product (a drug) rather than a function (long acting).[96]

The criteria for determining whether to include a potential respondent in the survey should be objective and clearly conveyed, preferably using written instructions addressed to those who administer the screening questions. These instructions and the completed screening questionnaire should be made available to the court and the opposing party along with the interview or survey form for each respondent. Computerized administration, described below, has allowed for this to be automated.

Survey Questions and Structure

Clarity, Precision, and Lack of Bias in the Framing of the Survey Questions

Although it seems obvious that questions on a survey should be clear and precise, phrasing questions to reach that goal is often difficult. Even questions that appear clear can convey unexpected meanings and ambiguities to potential respondents. For example, the question "What is the average number of days each week you have butter?" appears to be straightforward. Yet some respondents wondered whether margarine counted as butter, and when the question was revised to include the introductory phrase "not including margarine," the reported frequency of butter use dropped dramatically.[97]

95. 858 F. Supp. 1305, 1321 & n.13 (S.D.N.Y. 1994).

96. *Id.* at 1321.

97. Floyd J. Fowler, Jr., *How Unclear Terms Affect Survey Data*, 56 Pub. Op. Q. 218, 225–26 (1992), https://doi.org/10.1086/269312.

When unclear questions are included in a survey, they may threaten the validity of the survey by systematically distorting responses if respondents are misled in a particular direction, or by inflating random error if respondents guess because they do not understand the question.[98] If the crucial question is sufficiently ambiguous or unclear, it may be the basis for rejecting the survey. For example, a survey was designed to assess community sentiment that would warrant a change of venue in trying a case for damages sustained when a hotel skywalk collapsed.[99] The court found that the question "Based on what you have heard, read or seen, do you believe that in the current compensatory damage trials, the defendants, such as the contractors, designers, owners, and operators of the Hyatt Hotel, should be punished?" could neither be correctly understood nor easily answered.[100] The court noted that the phrase "compensatory damages," although well-defined for attorneys, was unlikely to be meaningful for laypersons.[101]

Pilot work is a standard and valuable way to improve the quality of a survey.[102] When there is any doubt whether a term or phrase will be clear to respondents, researchers should pretest to assess understanding, which may include focus groups or cognitive interviewing.[103]

The value of pilot work is greatest when the issues and questions in the survey differ from the issues and questions addressed in previous surveys, or when the target population differs significantly from previously surveyed populations.[104]

98. *See id*. at 219.

99. Firestone v. Crown Ctr. Redevelopment Corp., 693 S.W.2d 99 (Mo. 1985) (en banc).

100. *See id*. at 102, 103.

101. *See id*. at 103. When there is any question about whether some respondents will understand a particular term or phrase, the term or phrase should be defined explicitly in the survey.

102. *See* Jon A. Krosnick & Stanley Presser, *Questions and Questionnaire Design*, *in* Handbook of Survey Research, *supra* note 1, at 294 ("No matter how closely a questionnaire follows recommendations based on best practices, it is likely to benefit from pretesting. . . ."). *See also* Jean M. Converse & Stanley Presser, Survey Questions: Handcrafting the Standardized Questionnaire 51 (1986); Fred W. Morgan, *Judicial Standards for Survey Research: An Update and Guidelines*, 54 J. Mktg. 59, 64 (1990), https://doi.org/10.1177/002224299005400104; OMB Standards and Guidelines for Statistical Surveys, Standard 1.4, Pretesting Survey Systems (2006), https://perma.cc/D8XZ-4UX7 (specifying that to ensure that all components of a survey function as intended, pretests of survey components should be conducted unless those components have previously been successfully fielded); American Association for Public Opinion Research, Best Practices (2022), https://perma.cc/Y3HU-XCHK ("Before fielding a survey, it is important to pretest the questionnaire.").

103. Cognitive interviewing includes a combination of think-aloud and verbal probing techniques. Gordon B. Willis et al., *Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques*, *in* Cognition and Survey Research 136 (Monroe G. Sirken et al. eds., 1999). *See also* Gordon B. Willis, *Cognitive Interviewing in Survey Design: State of the Science and Future Directions*, *in* The Palgrave Handbook of Survey Research 103–07 (2018). *See also* Tourangeau et al., *supra* note 88, at 326–27.

104. Ivan R. Ross, *The Use of Pilot Tests and Pretests in Consumer Surveys*, *in* Trademark and Deceptive Advertising Surveys 13, 26 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

In many pretests or pilot tests,[105] the proposed survey is administered to a small sample (usually between twenty-five and seventy-five)[106] of the same type of respondents who would be eligible to participate in the full-scale survey. Some courts have explicitly recognized the value of pretests[107] and that lack of pretesting may suggest a weakness in the survey.[108]

Litigants would be more likely to conduct pilot work and disclose it in expert reports if courts recognized that pilot work can maximize the likelihood that respondents understand the questions they are being asked. Moreover, the Federal Rules of Civil Procedure may require that a testifying expert disclose pilot work that serves as a basis for the expert's opinion. The situation is more complicated when a nontestifying expert conducts the pilot work and the testifying expert learns about the pilot testing only indirectly through the attorney's advice about the relevant issues in the case. Some commentators suggest that attorneys are obligated to disclose such pilot work.[109]

What Respondents and Interviewers Knew About the Survey Purpose and Its Sponsorship

One way to protect the objectivity of survey administration is to avoid telling respondents or interviewers who is sponsoring the survey. Respondents who know the identity of the sponsor of the survey may adjust their responses, and interviewers who know the identity of the survey's sponsor may affect results inadvertently by communicating to respondents their expectations or what they believe are the preferred responses of the survey's sponsor. To ensure objectivity in the administration of the survey, it is standard interview practice in surveys conducted for litigation to do double-blind research whenever possible: Both the interviewer and the respondent are blind to the sponsor of the survey and its

105. The terms *pretest* and *pilot test* are sometimes used interchangeably to describe pilot work done in the planning stages of research. When they are distinguished, the difference is that a pretest tests the questionnaire, whereas a pilot test generally tests proposed collection procedures as well.

106. Converse & Presser, *supra* note 102, at 69. Converse and Presser suggest that a pretest with twenty-five respondents is appropriate when the survey uses professional interviewers.

107. *See, e.g.*, Zippo Mfg. Co. v. Rogers Imports, Inc., 216 F. Supp. 670 (S.D.N.Y. 1963); Scott v. City of New York, 591 F. Supp. 2d 554, 560 (S.D.N.Y. 2008) ("[T]he survey went through multiple pretests in order to insure its usefulness and statistical validity."); Estes Park Taffy Co. v. Original Taffy Shop, Inc., No. 15-cv-01697-CBS, 2017 U.S. Dist. LEXIS 88113, at *9 (D. Colo. June 8, 2017).

108. GOLO, LLC. v. Goli Nutrition, Inc., No.20-667-RGA, 2020 U.S. Dist. LEXIS 158508, at *28 (D. Del. Sept. 1, 2020) (including lack of pretesting in a list of potential weaknesses in an expert's method).

109. *See* Yvonne C. Schroeder, *Pretesting Survey Questions: The Procedural and Ethical Ramifications*, 11 Am. J. Trial Advoc. 195, 197–201 (1987).

purpose. Thus, the survey instrument provides no explicit or implicit clues about the sponsorship of the survey (e.g., a sponsor's letterhead) or the expected responses (e.g., reversing the usual order of the yes and no response boxes on a key question, potentially increasing the likelihood that *no* will be checked).[110]

Nonetheless, in some situations (e.g., on some government surveys), disclosure of the survey's sponsor to respondents (and thus to interviewers) is required. Such surveys call for an evaluation of the likely biases introduced by interviewer or respondent awareness of the survey's sponsorship. In evaluating the consequences of sponsorship awareness, it is important to consider (1) whether the sponsor has views and expectations that are apparent and (2) whether awareness is confined to the interviewers or involves the respondents. For example, if a survey concerning attitudes toward gun control is sponsored by the National Rifle Association, it is clear that responses opposing gun control are likely to be preferred. In contrast, if the survey on gun control attitudes is sponsored by the Department of Justice, the identity of the sponsor may not suggest the kinds of responses the sponsor expects or would find acceptable.[111] When a survey involves interviewers who are well-trained, their awareness of sponsorship may be a less serious threat than respondents' awareness.[112]

In most situations, the survey expert can follow the traditional CASRO Code of Standards and Ethics[113] and promise the respondent anonymity in the interest of obtaining the most accurate and candid responses. Moreover, in most situations, there is no need to disclose the identity of the survey sponsor. In some cases, however, respondents to a survey are involved in litigation (e.g., class-action wage and hour cases) and are likely to recognize that a survey asking them questions about relevant issues (such as unpaid overtime) is being sponsored by a party to the litigation. They may even be unwilling to participate in the survey unless they are told that the attorney representing them is sponsoring it. This creates a difficult problem for survey researchers. To get an acceptable response rate, it may be necessary to reveal who is sponsoring the survey. But this revelation may incentivize respondents to adjust their answers for potential financial gain, and promising anonymity may exacerbate this issue. If the survey is conducted by a friendly party, one approach that can be used to bolster accuracy is to tell respondents

110. *See* Centaur Commc'ns, Ltd. v. A/S/M Commc'ns, Inc., 652 F. Supp. 1105, 1111 n.3 (S.D.N.Y.) (pointing out that reversing the usual order of response choices, "yes or no," to "no or yes" may confuse interviewers as well as introduce bias), *aff'd*, 830 F.2d 1217 (2d Cir. 1987).

111. *See, e.g.*, Stanley Presser et al., *Survey Sponsorship, Response Rates, and Response Effects*, 73 Soc. Sci. Q. 699, 701 (1992) (different responses to a university-sponsored telephone survey and a newspaper-sponsored survey for questions concerning attitudes toward the mayoral primary, an issue on which the newspaper had taken a position).

112. *See, e.g.*, Seymour Sudman et al., *Modest Expectations: The Effects of Interviewers' Prior Expectations on Responses*, 6 Soc. Methods & Rsch. 171, 181 (1977), https://doi.org/10.1177/004912417700600203.

113. CASRO, *supra* note 41, § I.A.

that their responses may not be anonymous and thus that exaggeration or other inaccuracies may be detected. It is unclear what the best practice is in this case, and the survey expert should be prepared to justify their choices.

Handling Respondents with No Opinions and Reducing Respondent Guessing

Some survey respondents may have no opinion about an issue under investigation, either because they have never thought about it before or because the question mistakenly assumes a familiarity with the issue. For example, some respondents in a consumer survey may not have noticed that the commercial they are being questioned about guaranteed the quality of the product being advertised, and thus they may have no opinion on the kind of guarantee it indicated. The following three alternative question structures will affect how those respondents answer and how their responses are counted.

First, the survey can ask all respondents to answer the question (e.g., "Did you understand the guarantee offered by Clover to be a 1-year guarantee, a 60-day guarantee, or a 30-day guarantee?"). Faced with a direct question, particularly one that provides response alternatives, the respondent obligingly may supply an answer even if (in this example) the respondent did not notice the guarantee. Such answers will reflect only what the respondent can glean from the question, or they may reflect pure guessing.

Second, the survey can use a quasi-filter question to reduce guessing by providing "don't know" or "no opinion" options as part of the question (e.g., "Did you understand the guarantee offered by Clover to be for more than a year, a year, or less than a year, or don't you have an opinion?").[114] By signaling to the respondent that it is acceptable not to have an opinion, the question reduces the demand for an answer and, as a result, the inclination to hazard a guess just to comply. Respondents are more likely to choose a "no opinion" option if it is mentioned explicitly by an interviewer than if it is merely accepted when the respondent spontaneously offers it as a response. Similarly, in an online survey, explicitly providing "don't know" as a potential response rather than accepting it only if the respondent writes it in as an "other" response can increase the use of that option. The consequence of this change in format can be substantial. Studies indicate that, although the relative distribution of the respondents selecting the *listed* choices is unlikely to change dramatically, presentation of an explicit "don't

114. Norbert Schwarz & Hans-Jürgen Hippler, *Response Alternatives: The Impact of Their Choice and Presentation Order, in* Measurement Errors in Surveys 41, 45–46 (Paul P. Biemer et al. eds., 1991); Spangler Candy Co. v. Tootsie Roll Indus., LLC, 372 F. Supp. 3d 588, 599 (N.D. Ohio 2019) ("Because [the expert] offered a 'Don't Know/No Opinion' option for each closed-ended question, I conclude the questions were not unduly suggestive or guess-inducing.").

know" or "no opinion" alternative commonly leads to a 20% to 25% increase in the proportion of respondents selecting that response.[115]

Finally, the survey can include full-filter questions—that is, questions that lay the groundwork for the substantive question by first asking respondents if they have an opinion about the issue or happened to notice the feature that the interviewer is preparing to ask about (e.g., "Based on the commercial you just saw, do you have an opinion about how long Clover stated or implied that its guarantee lasts?").[116] The survey then asks the substantive question only of those respondents who have indicated that they have an opinion on the issue.

Which of these three approaches is used and the way it is used can affect the rate of "no opinion" responses that the substantive question will evoke.[117] Respondents are more likely to say that they do not have an opinion on an issue if a full filter is used than if a quasi-filter is used.[118] However, in maximizing respondent expressions of "no opinion," full filters may produce an underreporting of opinions. Some evidence indicates that full-filter questions discourage respondents who actually have opinions from offering them by conveying the implicit suggestion that respondents can avoid difficult follow-up questions by saying that they have no opinion.[119]

In general, then, a survey that uses full filters provides a conservative estimate of the number of respondents holding an opinion, while a survey that uses neither full filters nor quasi-filters may overestimate the number of respondents with opinions if some respondents offering opinions are guessing. The strategy of including a "no opinion" or "don't know" response as a quasi-filter avoids both of these extremes, although some research suggests that even a quasi-filter may discourage a substantive answer from a respondent who would be able to provide one.[120]

One solution that some survey researchers use is to provide respondents with a general instruction not to guess at the beginning of an interview, rather than supplying a "don't know" or "no opinion" option as part of the options attached to each question.[121] Another approach is to eliminate the "don't know" option and to add follow-up questions that measure the strength of the respondent's

115. Howard Schuman & Stanley Presser, Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context 113–46 (1996).

116. *See, e.g.*, Johnson & Johnson * Merck Consumer Pharms. Co. v. SmithKline Beecham Corp., 960 F.2d 294, 299 (2d Cir. 1992).

117. Considerable research has been conducted on the effects of filters. For a review, see George F. Bishop et al., *Effects of Filter Questions in Public Opinion Surveys*, 47 Pub. Op. Q. 528 (1983), https://doi.org/10.1086/268810.

118. Schwarz & Hippler, *supra* note 114, at 45–46.

119. *Id.* at 46.

120. Jon A. Krosnick et al., *The Impact of "No Opinion" Response Options on Data Quality: Non-Attitude Reduction or Invitation to Satisfice?*, 66 Pub. Op. Q. 371 (2002), https://doi.org/10.1086/341394.

121. Anheuser-Busch, Inc. v. VIP Prods., LLC, No. 4:08cv0358, 2008 U.S. Dist. LEXIS 82258, at *6 (E.D. Mo. Oct. 16, 2008).

opinion.[122] More generally, survey experts can conduct pilot tests or cognitive interviews to assess whether many in the population in fact lack any opinion or the relevant knowledge to arrive at an opinion. If a substantial majority of people do have an opinion, it may be appropriate to exclude a "don't know" option and thus avoid reducing data quality by encouraging less motivated respondents to use that response. In contrast, if many people do not have formed opinions or the knowledge required to arrive at them, the option should be included.[123] The survey expert should be prepared to explain why the "don't know" option was included or excluded.

Use of Open-Ended and Closed-Ended Questions

The questions that make up a survey instrument may be open-ended, closed-ended, or a combination of both. Both are valid choices in appropriate situations. Open-ended questions require the respondent to formulate and express an answer in their own words (e.g., "What was the main point of the commercial?"). Closed-ended questions provide the respondent with an explicit set of responses from which to choose; the choices may be as simple as *yes* or *no* (e.g., "Is Colby College coeducational?"[124]) or as complex as a range of alternatives (e.g., "The two pain relievers have (1) the same likelihood of causing gastric ulcers; (2) about the same likelihood of causing gastric ulcers; (3) a somewhat different likelihood of causing gastric ulcers; (4) a very different likelihood of causing gastric ulcers; or (5) none of the above."[125]). When a survey involves in-person interviews, the interviewer may show the respondent these choices on a showcard that lists them.

Open-ended and closed-ended questions may elicit very different responses.[126] Most responses are less likely to be volunteered by respondents who are asked an

122. Krosnick & Presser, *supra* note 102, at 285.

123. Dragana Bolcic-Jankovic et al., *Using "Don't Know" Responses in a Survey of Oncologists Regarding Medicinal Cannabis*, 14 Surv. Prac. (2021), https://doi.org/10.29115/SP-2020-0016; *see also* Erika A. Waters et al., *Dismissing "Don't Know" Responses to Perceived Risk Survey Items Threatens the Validity of Theoretical and Empirical Behavior-Change Research*, 17 Persps. Psych. Sci. 841 (2022), https://doi.org/10.1177/17456916211017860.

124. President & Trs. of Colby College v. Colby College–N.H., 508 F.2d 804, 809 (1st Cir. 1975).

125. This question is based on one asked in *American Home Products Corp. v. Johnson & Johnson*, 654 F. Supp. 568, 581 (S.D.N.Y. 1987), that was found to be a leading question by the court, primarily because the choices suggested that the respondent had learned about aspirin's and ibuprofen's relative likelihoods of causing gastric ulcers. In contrast, in *McNeilab, Inc. v. American Home Products Corp.*, 501 F. Supp. 517, 525 (S.D.N.Y. 1980), the court accepted as nonleading the question, "Based only on what the commercial said, would Maximum Strength Anacin contain more pain reliever, the same amount of pain reliever, or less pain reliever than the brand you, yourself, currently use most often?"

126. Howard Schuman & Stanley Presser, *Question Wording as an Independent Variable in Survey Analysis*, 6 Socio. Methods & Rsch. 151 (1977), https://doi.org/10.1177/004912417700600202; Schuman & Presser, *supra* note 115, at 79–112; Converse & Presser, *supra* note 102, at 33.

open-ended question than they are to be chosen by respondents who are presented with a closed-ended question. The response alternatives in a closed-ended question may remind respondents of options that they would not otherwise consider or which simply do not come to mind as easily.[127]

The advantage of open-ended questions is that they give the respondent fewer hints about expected or preferred answers. Response choices in a closed-ended question, in addition to reminding respondents of options that they might not otherwise consider, may direct the respondent away from or toward a particular response. For example, a commercial reported that in shampoo tests with more than 900 women, the sponsor's product received higher ratings than other brands.[128] According to a competitor, the commercial deceptively implied that each woman in the test rated more than one shampoo, when in fact each woman rated only one. To test consumer impressions, the survey showed the commercial and asked: "How many different brands mentioned in the commercial did each of the 900 women try?"[129] Respondents were given the choice of "one," "two," "three," "four," or "five or more." The fact that four of the five choices in the closed-ended question provided a response that was greater than one implied that the correct answer was probably more than one.[130]

An open-ended question may also suggest that the answer is more than one, however. By asking "how many different brands," the question suggests (1) that the viewer should have received some message from the commercial about the number of brands each woman tried, and (2) that different brands were tried. Instead, a nonleading version of the question would have simply asked: "How many brands mentioned in the commercial did each of the 900 women try?"

127. For example, when respondents in one survey were asked, "What is the most important thing for children to learn to prepare them for life?", 62% picked "to think for themselves" from a list of five options, but only 5% spontaneously offered that answer when the question was open-ended. Schuman & Presser, *supra* note 115, at 104–07. An open-ended question presents the respondent with a free-recall task, whereas a closed-ended question is a recognition task. Recognition tasks in general reveal higher performance levels than recall tasks. Mary M. Smyth et al., Cognition in Action 25 (1987). In addition, there is evidence that respondents answering open-ended questions may be less likely to report some information that they would reveal in response to a closed-ended question when that information seems self-evident or irrelevant.

128. *See* Vidal Sassoon, Inc. v. Bristol-Myers Co., 661 F.2d 272, 273 (2d Cir. 1981).

129. This was the wording of the closed-ended question in the survey discussed in *Vidal Sassoon*, 661 F.2d at 275–76.

130. Ninety-five percent of the respondents who answered the closed-ended question in the plaintiff's survey said that each woman had tried two or more brands. The open-ended question was never asked. *Vidal Sassoon*, 661 F.2d at 276. Norbert Schwarz, *Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction*, *in* Research Methods in Personality and Social Psychology 98 (Clyde Hendrick & Margaret S. Clark eds., 1990), suggests that respondents often rely on the range of response alternatives as a frame of reference when they are asked for frequency judgments. *See, e.g.*, Roger Tourangeau & Tom W. Smith, *Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context*, 60 Pub. Op. Q. 275, 292 (1996), https://doi.org/10.1086/297751.

rather than "How many *different* brands" did each try? Similarly, an open-ended question that asks, "[W]hich company or store do you think puts out this shirt?" indicates to the respondent that the appropriate answer is the name of a company or store. The question would be leading if the respondent would have considered other possibilities (e.g., an individual or website if the question had not provided the frame of a company or store).[131] Thus, the wording of a question, open-ended or closed-ended, can be leading, and the degree of suggestiveness of each question must be considered in evaluating the objectivity of a survey.

Closed-ended questions have some additional potential weaknesses that arise if the choices are not constructed properly. If the respondent is asked to choose one or more responses from among several choices, the response or responses chosen will be meaningful only if the list of choices is exhaustive—that is, if the choices cover all possible answers a respondent might give to the question. If the list of possible choices is incomplete, a respondent may be forced to choose one that does not express their opinion.[132] The omission of a relevant choice option can only be partially attenuated by telling respondents explicitly that they are not limited to the choices presented, because most respondents nevertheless will select an answer from among the listed ones.[133]

One form of closed-ended question format that can produce some distortion is the popular agree/disagree, true/false, or yes/no question. Although this format is appealing because it is easy to write and score these questions and their responses, the format is also seriously problematic. With its simplicity comes acquiescence: "[T]he tendency to endorse any assertion made in a question, regardless of its content," is a systematic source of bias that has produced an inflation effect of 10% across a number of studies.[134] Only when control groups or control questions are added to the survey design, or when multiple items are counterbalanced, can this question format provide reasonable response estimates.[135]

One challenge with open-ended responses is translating them into meaningful numeric metrics.[136] Imagine that a researcher asks respondents what they think about self-driving cars. If the researcher were interested in the prevalence and nature of safety concerns about self-driving cars, they would need a mechanism for translating the varied open-ended responses into a set of binary values: each possible safety concern was listed or not. This requires developing a list of different types of concerns such that each concern is unique (e.g., can malfunction, inability of driver to have control, etc.), and each has a text label. The list

131. Smith v. Wal-Mart Stores, Inc., 537 F. Supp. 2d 1302, 1331–32 (N.D. Ga. 2008).

132. *See, e.g.*, Am. Home Prods. Corp. v. Johnson & Johnson, 654 F. Supp. 568, 581 (S.D.N.Y. 1987).

133. *See* Howard Schuman, *Ordinary Questions, Survey Questions, and Policy Questions*, 50 Pub. Op. Q. 432, 435–36 (1986), https://doi.org/10.1093/pubopq/50.3.432.

134. Krosnick, *supra* note 86, at 552–53.

135. *See* section titled "Potential Order Effects" below.

136. Groves et al., *supra* note 77, at 332–44.

should also be exhaustive (cover all written answers) and include both a code for responses that do not refer to safety issues (e.g., reflects technological evolution) and also a code for nonresponses (e.g., left blank). In some cases, a respondent may give more than one answer, and that response should be coded in more than one category (e.g., "can malfunction and has no human control"). The researcher should be transparent on the construction of the coding system, which may be based on a priori categories or constructed based on the answers the respondents have given. Moreover, unless there are substantial data privacy concerns, the full open-ended answers should be made available to the other party for analysis.[137]

The coding system should be clear enough that different coders would assign the same responses to the same categories in the vast majority of cases. In most academic research, this is accomplished by having multiple coders who are blind to the hypotheses and purpose of the project, giving them detailed instructions and training, and measuring inter coder reliability. Although this approach constitutes the best way to ensure reliable coding, because the raw data underlying the coding and the coding itself are made available to the court and the opposing party in litigation, the court and opposing party can directly examine and evaluate the appropriateness of the coding. It is therefore not uncommon for coding of simple measures in litigation to dispense with having multiple blind coders (though the resultant coding should be given close scrutiny).[138]

Although many courts prefer open-ended questions on the ground that they are likely to be less leading, the value of any open-ended or closed-ended question depends on the information it conveys in the question and, in the case of closed-ended questions, in the choices provided. Open-ended questions are more appropriate when the survey is attempting to gauge what comes first to a respondent's mind, but closed-ended questions are more suitable for assessing choices between well-identified options or obtaining ratings on a clear set of alternatives.

137. Mary B. Vardigan & Peter Granda, *Archiving, Documentation, and Dissemination*, *in* Handbook of Survey Research 707 (Peter V. Marsden & James D. Wright eds., 2d ed. 2010); *see also* section titled "Disclosure and Reporting" *infra*. Partial redaction of a response might be appropriate if the response was individually identifiable and the content sensitive. For example, a complaint about a named supervisor in an employment survey would fall in this category.

138. *See, e.g.*, Revlon Consumer Prods. Corp. v. Jennifer Leather Broadway, Inc., 858 F. Supp. 1268, 1276 (S.D.N.Y. 1994) (inconsistent scoring and subjective coding led court to find survey so unreliable that it was entitled to no weight), *aff'd*, 57 F.3d 1062 (2d Cir. 1995); Rock v. Zimmerman, 959 F.2d 1237, 1253 n.9 (3d Cir. 1992) (court found that responses on a change-of-venue survey incorrectly categorized respondents who believed the defendant was insane as believing he was guilty); Coca-Cola Co. v. Tropicana Prods., Inc., 538 F. Supp. 1091, 1094–96 (S.D.N.Y.) (plaintiff's expert stated that respondents' answers to the open-ended questions revealed that 43% of respondents thought Tropicana was portrayed as fresh-squeezed; the court's own tabulation found no more than 15% believed this was true), *rev'd on other grounds*, 690 F.2d 312 (2d Cir. 1982); *see also* Cumberland Packing Corp. v. Monsanto Co., 140 F. Supp. 2d 241 (E.D.N.Y. 2001) (court examined verbatim responses that respondents gave to arrive at a confusion level substantially lower than the level reported by the survey expert).

Use of Probes to Clarify Ambiguous or Incomplete Answers

When questions allow respondents to express their opinions in their own words, some of the respondents may give ambiguous or incomplete answers. If the survey is administered online, some probes (e.g., "Why did you choose that answer?") can be programmed into the survey instrument and automatically administered as part of the survey. If interviewers are asking the questions, they may be instructed to probe to obtain a more complete response or clarify the meaning of the ambiguous response. Interviewers should be instructed what clarification, if any, they can provide. The record should reflect both what the respondent said initially, what the interviewer said in the attempt to get or provide clarification, and how the respondent answered the probe; this information will allow the court and the opposing party to evaluate whether the probe affected the views expressed by the respondent.

If the survey involves interviewers who are permitted to administer follow-up questions, they must be given explicit instructions on when they should probe and what they should say in probing.[139] Standard probes used to draw out all that the respondent has to say (e.g., "Any further thoughts?"; "Anything else?"; "Can you explain that a little more?"; or "Could you say that another way?") are relatively noncontroversial and can be programmed into an online survey instrument. But probing should be limited. Persistent continued requests for further responses to the same or nearly identical questions may convey the idea to the respondent that they have not yet produced the "right" answer, particularly if the probes are administered by a live interviewer.[140]

Potential Order Effects

The order in which questions are asked on a survey and the order in which response alternatives are provided in a closed-ended question can influence the answers.[141] For example, although asking a general question before a more specific

139. Floyd J. Fowler, Jr. & Thomas W. Mangione, Standardized Survey Interviewing: Minimizing Interviewer-Related Error 41–42 (1990), https://doi.org/10.4135/9781412985925.

140. *See, e.g.*, Johnson & Johnson-Merck Consumer Pharms. Co. v. Rhone-Poulenc Rorer Pharms., Inc., 19 F.3d 125, 135 (3d Cir. 1994); Am. Home Prods. Corp. v. Procter & Gamble Co., 871 F. Supp. 739, 748 (D.N.J. 1994).

141. *See* Schuman & Presser, *supra* note 115, at 23, 56–74; Krosnick & Presser, *supra* note 102, at 278–81. In *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*, 511 F. Supp. 867, 875 (S.D.N.Y. 1980), the court recognized the biased structure of a survey that disclosed the tar content of the cigarettes being compared before questioning respondents about their cigarette preferences. Not surprisingly, respondents expressed a preference for the lower tar product. *See also* E. & J. Gallo Winery v. Pasatiempos Gallo, S.A., 905 F. Supp. 1403, 1409–10 (E.D. Cal. 1994) (court recognized

question on the same topic is unlikely to affect the response to the specific question, reversing the order of the questions may influence responses to the general question. As a rule, then, surveys are less likely to be subject to order effects if the questions move from the general (e.g., "What do you recall being discussed in the advertisement?") to the specific (e.g., "Based on your reading of the advertisement, what companies do you think the ad is referring to when it talks about rental trucks that average five miles per gallon?").[142]

The mode of questioning can influence the form that an order effect takes. When respondents are shown response alternatives visually, as in mail surveys and self-administered online surveys or in face-to-face interviews when respondents are shown a card containing response alternatives, they are more likely to select the first choice offered (a primacy effect).[143] In contrast, when response alternatives are presented orally, as in telephone surveys, respondents are more likely to choose the last choice offered (a recency effect).[144] Although these effects are typically small, no general formula is available that can adjust values to correct for order effects, because the size and even the direction of the order effects may depend on the nature of the question being asked and the choices being offered. To control for order effects, the order of the response choices in a survey should be rotated, randomized, or counterbalanced,[145] so that no response alternative will have an inflated chance of being selected because of its position.

Appropriate Inclusion of Control Groups, Control Questions, or Other Comparisons

Many surveys are designed not simply to describe attitudes, beliefs, or reported behaviors, but to determine the source of those attitudes, beliefs, or behaviors. That is, the purpose of the survey is to test a causal proposition. For example, how does a trademark or the content of a commercial affect respondents' perceptions or understanding of a product or commercial? Thus, the question is not merely whether consumers hold inaccurate beliefs about Product A, but whether exposure to the commercial misleads the consumer into thinking that Product A

that earlier questions referring to playing cards, board or table games, or party supplies, such as confetti, increased the likelihood that respondents would include these items in answers to the questions that followed).

142. This question was accepted by the court in *U-Haul International, Inc. v. Jartran, Inc.*, 522 F. Supp. 1238, 1249 (D. Ariz. 1981), *aff'd*, 681 F.2d 1159 (9th Cir. 1982).

143. Krosnick & Presser, *supra* note 102, at 280.

144. *Id.*

145. *See, e.g.*, Winning Ways, Inc. v. Holloway Sportswear, Inc., 913 F. Supp. 1454, 1465–67 (D. Kan. 1996) (failure to rotate the order in which the jackets were shown to the consumers led to reduced weight for the survey); Procter & Gamble Pharms., Inc. v. Hoffmann-La Roche, Inc., No. 06 Civ. 0034 (PAC), 2006 U.S. Dist. LEXIS 64363 (S.D.N.Y. Sept. 6, 2006).

is a superior pain reliever. Yet if consumers already believe, before viewing the commercial, that Product A is a superior pain reliever, a survey that simply records consumers' impressions after they view the commercial may reflect those preexisting beliefs rather than impressions produced by the commercial.

Some surveys attempt to reduce the impact of preexisting impressions on respondents' answers by instructing respondents to focus solely on the stimulus as a basis for their answers. Thus, the survey includes a preface (e.g., "based on the commercial you just saw") or directs the respondent's attention to the mark at issue (e.g., "these stripes on the package"). Such efforts are likely to be only partially successful. It is often difficult for respondents to identify accurately the source of their impressions.[146] The more routine the idea being examined in the survey, the more likely it is that the respondent's answer is influenced by (1) preexisting impressions; (2) general expectations about what commercials typically say (e.g., the product being advertised is better than its competitors); or (3) guessing, rather than by the actual content of the commercial message or the trademark being evaluated. A similar limitation occurs when respondents are asked to explain why they chose a particular response: they can justify their choice, but may not be able to report accurately what actually led them to make that choice.

With respondents randomly assigned to one or more appropriate comparison conditions, the survey expert can draw clear causal inferences about the influence of the stimulus.[147] In the simplest version of such a survey-experiment (as discussed above), respondents are assigned randomly to one of two conditions.[148] For example, respondents assigned to the experimental condition view an allegedly deceptive commercial, and respondents assigned to the control condition view the same commercial with the allegedly deceptive material removed or an alternative commercial that does not contain the allegedly deceptive material.[149]

146. *See* Richard E. Nisbett & Timothy D. Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 Psych. Rev. 231 (1977), https://doi.org/10.1037/0033-295X.84.3.231.

147. *See* Shari S. Diamond, *Using Psychology to Control Law: From Deceptive Advertising to Criminal Sentencing*, 13 L. & Hum. Behavior 239, 244–46 (1989), https://doi.org/10.1007/BF01067028; Jacob Jacoby & Constance Small, *Applied Marketing: The FDA Approach to Defining Misleading Advertising*, 39 J. Mktg. 65, 68 (1975), https://doi.org/10.1177/002224297503900413. *See also* R. Charles Henn, *Why Ask Why? A Critical Assessment of an Historical Survey Artifact*, 113 Trademark Rep. 772, 773–76 (2023).

148. Random assignment should not be confused with random selection. When respondents are assigned randomly to different treatment groups (e.g., respondents in each group watch a different commercial), the procedure ensures that within the limits of sampling error the two groups of respondents will be equivalent, on average, except for the different treatments they receive. Respondents selected for a mall intercept study, and not from a probability sample, may be assigned randomly to different treatment groups. Random selection, in contrast, describes the method of selecting a sample of respondents in a probability sample. *See* section titled "The Sample as a Reflection of the Relevant Characteristics of the Population" above.

149. This alternative commercial could be a "tombstone" advertisement that includes only the name of the product or a more elaborate commercial that does not include the claim at issue.

Respondents in both the experimental and control groups answer the same set of questions about the allegedly deceptive message. The effect of the commercial's allegedly deceptive message is evaluated by comparing the responses made by the experimental group members with those of the control group members. If 40% of the respondents in the experimental group responded indicating a belief in the deceptive claim, whereas only 8% of the respondents in the control group gave that response, the difference between 40% and 8% (within the limits of sampling error[150]) can be attributed to the allegedly deceptive commercial.

A survey-experimental design with an appropriate control group can account for more than preexisting beliefs.[151] Other sources of systematic and random error can influence responses to survey questions. Thus, if a respondent is asked "Have you heard of a security company named Titan Alarm?" some individuals will say they have, even if they have not. A control group can help address this type of error as well; this and other background noise should have produced similar response levels in the experimental and control groups, so comparing average scores in those groups should control for those sources of error. In addition, a leading question may cause participants to be more likely to endorse or reject a particular statement. But if respondents who viewed the allegedly deceptive commercial respond differently than respondents who viewed the control commercial, the difference cannot be merely the result of a leading question, because both groups answered the same question. The ability to evaluate the effect of the wording of a particular question makes the control group design particularly useful in assessing responses to closed-ended questions,[152] which may encourage guessing or particular responses. Thus, the focus is not on the absolute response level in the experimental group, but on the difference between the responses from the experimental group and those from the control group.[153]

In designing a survey-experiment, the expert should select a stimulus for the control group that shares as many characteristics with the experimental stimulus as possible, with the key exception of the characteristic whose influence is being assessed.[154] Although a survey with an imperfect control group may provide

150. For a discussion of sampling error, see the glossary to the current chapter and David H. Kaye and Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

151. For a more extensive discussion of controls, see Shari Seidman Diamond, *Control Foundations: Rationale and Approaches*, *in* Trademark and Deceptive Advertising Surveys 239 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

152. The Federal Trade Commission has long recognized the need for some kind of control for closed-ended questions, although it has not specified the type of control that is necessary. *See In re* Stouffer Foods Corp., 118 F.T.C. 746, 808–09 (Sept. 26, 1994).

153. *See, e.g.*, CytoSport, Inc. v. Vital Pharms., Inc., 617 F. Supp. 2d 1051, 1075–76 (E.D. Cal. 2009) (net confusion level of 25.4% obtained by subtracting 26.5% in the control group from 51.9% in the test group).

154. *See, e.g.*, Skechers USA, Inc. v. Vans, Inc., No. CV-07-01703 DSF (PLAx), 2007 WL 4181677, at *8–9 (C.D. Cal. Nov. 20, 2007) (in trade dress infringement case, control stimulus should have retained design elements not at issue); Procter & Gamble Pharms., Inc. v. Hoffman-LaRoche,

better information than a survey with no control group at all, the choice of an appropriate control group requires care and should influence the admissibility of the survey and the weight that the survey receives. For example, a control stimulus should not be less attractive than the experimental stimulus if the survey is designed to measure how familiar the experimental stimulus is to respondents, because attractiveness may affect perceived familiarity.[155] Nor should the control stimulus share with the experimental stimulus the feature whose impact is being assessed. If the control stimulus in a case of alleged trademark infringement is itself a likely source of consumer confusion, for example, reactions to the experimental and control stimuli may not differ because both cause respondents to express a similar level of confusion.[156] In an extreme case, an inappropriate control may do nothing more than control for the effect of the nature or wording of the survey questions (e.g., acquiescence).[157] That generally will not be enough to rule out other explanations for different or similar responses to the experimental and control stimuli. Finally, it may sometimes be appropriate to have more than one control group to assess precisely what is causing the response to the experimental stimulus (e.g., in the case of an allegedly deceptive ad, whether it is a misleading graph or a misleading claim by the announcer, or in the case of allegedly infringing trade dress, whether it is the style of the font used or the coloring of the packaging).[158]

Courts have increasingly come to recognize the central role that control groups play in evaluating causal claims.[159] Litigants have taken the cue, and most

Inc., No. 06 Civ 0034 (PAC), 2006 U.S. Dist. LEXIS 64363, at *87–88 (S.D.N.Y. Sept. 6, 2006) (in false advertising action, disclaimer was inadequate substitute for appropriate control group).

155. *See, e.g.*, Indianapolis Colts, Inc. v. Metro. Balt. Football Club L.P., 34 F.3d 410, 415–16 (7th Cir. 1994) (court recognized that the name "Baltimore Horses" was less attractive for a sports team than the name "Baltimore Colts"); *see also* Reed-Union Corp. v. Turtle Wax, Inc., 77 F.3d 909, 912 (7th Cir. 1996) (court noted that one expert's choice of a control brand with a well-known corporate source was less appropriate than the opposing expert's choice of a control brand whose name did not indicate a specific corporate source).

156. *See, e.g.*, W. Publ'g Co. v. Publ'ns Int'l, Ltd., No. 94-C-6803, 1995 U.S. Dist. LEXIS 5917, at *45 (N.D. Ill. May 2, 1995) (court noted that the control product was "arguably more infringing than" the defendant's product) (emphasis omitted). *See also* Classic Foods Int'l Corp. v. Kettle Foods, Inc., No. SACV 04–725 CJC (Ex), 2006 U.S. Dist. LEXIS 97200 (C.D. Cal. Mar. 2, 2006); McNeil-PPC, Inc. v. Merisant Co., No. 04–1090 (JAG), 2004 U.S. Dist. LEXIS 27733 (D.P.R. July 29, 2004).

157. *See supra* text accompanying note 134.

158. *See, e.g.*, Masterfoods USA v. Arcor USA, Inc., 230 F. Supp. 2d 302 (W.D.N.Y. 2002).

159. *See, e.g.*, Colangelo v. Champion Petfoods USA, Inc., No. 6:18-CV-1228, 2022 U.S. Dist. LEXIS 60489, at *32 (N.D.N.Y. Mar. 31, 2022) ("[W]ithout a control group it is impossible to establish cause and effect."); Longoria v. Million Dollar Corp., No. 18-CV-02266-PAB-NYW, 2021 U.S.Dist. LEXIS 38478, at *27 (D. Colo. Mar. 2, 2021) (survey excluded because "a causal study . . . requires a control group"); SmithKline Beecham Consumer Healthcare, L.P. v. Johnson & Johnson-Merck, No. 01 Civ. 2775 (DAB), 2001 U.S. Dist. LEXIS 7061, at *37–39 (S.D.N.Y. June 1, 2001) (survey to assess implied falsity of a commercial not probative in the absence of a control group);

experts recognize the need to submit a survey with a control group (i.e., a survey-experiment) if they wish to make a causal claim and rule out other explanations for the responses to the survey questions.[160]

A less common control methodology is a control question. Rather than administering a control stimulus to a separate group of respondents, the survey asks all respondents one or more control questions along with the question about the product or service at issue. This technique is used to evaluate whether a brand name is generic. The genericness survey presents survey respondents with a series of product or service names and asks them to indicate in each instance whether they believe the name is a brand name or a common name. By showing that 68% of respondents considered Teflon a brand name (a proportion similar to the 75% of respondents who recognized the acknowledged trademark Jell-O as a brand name, and markedly different from the 13% who thought aspirin was a brand name), the makers of Teflon demonstrated that their respondents understood the difference between brands and product categories (so-called common names) and that the Teflon mark at issue was perceived as a brand name; they retained their trademark.[161] It is crucial to control for order effects in such designs.

The issue of appropriate comparisons is especially salient in the context of conjoint analysis, which has been used in some patent and false-advertising cases. In cases of patent infringement, the plaintiff must submit an estimate of the damages produced by the defendant's infringement. When the infringement involves a multicomponent product, the plaintiff must apportion damages to determine the value of the patented component at issue.[162] Similarly, in class actions involving false advertising, damages depend on the value to consumers that can be attributed to the deceptive portion of the advertisement.[163] To estimate these damages, survey experts have increasingly turned to conjoint analysis. When done

Am. Home Prods. Corp. v. Procter & Gamble Co., 871 F. Supp. 739, 749 (D.N.J. 1994) (discounting survey results based on failure to control for participants' preconceived notions); ConAgra, Inc. v. Geo. A. Hormel & Co., 784 F. Supp. 700, 728 (D. Neb. 1992) ("Since no control was used, the . . . study, standing alone, must be significantly discounted."), *aff'd*, 990 F.2d 368 (8th Cir. 1993).

160. William, R. Shadish et al., Experimental and Quasi-Experimental Designs for Generalized Causal Inference (2002); James N. Druckman, Experimental Thinking: A Primer on Social Science Experiments (2022).

161. E.I. DuPont de Nemours & Co. v. Yoshida Int'l, Inc., 393 F. Supp. 502, 526–27 & n.54 (E.D.N.Y. 1975); *see also* Donchez v. Coors Brewing Co., 392 F.3d 1211, 1218 (10th Cir. 2004) (respondents evaluated eight brand and generic names in addition to the disputed name); Reinalt-Thomas Corp. v. Mavis Tire Supply, LLC, 391 F. Supp. 3d 1261, 1271–73 (N.D. Ga. 2019). A similar approach is used in assessing secondary meaning. *See, e.g.*, T-Mobile US, Inc. v. AIO Wireless LLC, 991 F. Supp. 2d 888 (S.D. Tex. 2014).

162. For example, in *Apple, Inc. v. Samsung Electronics Co., Ltd.*, No. 11-CV-01846-LHK, 2014 U.S. Dist. LEXIS 29721 (N.D. Cal. Mar. 6, 2014), Apple used a conjoint survey to isolate the costs associated with Samsung's alleged patent infringement regarding iPhone and iPad features.

163. *E.g.*, Price v. L'Oréal U.S., Inc., No. 17-Civ-614 (LGS), 2020 Dist. LEXIS 153255 (S.D.N.Y. Aug. 24, 2020).

well, conjoint analysis can provide useful information, but there are many decisions made in designing a conjoint analysis that can undermine reliability and bias the resulting estimates.[164]

In contrast to a survey that asks respondents directly how much they would be willing to pay (WTP) for a given feature or attribute of a product, a task that may be difficult for consumers, a conjoint survey-experiment typically presents respondents with choices between products with different randomly assigned profiles consisting of combinations of features.[165] Suppose the survey is designed to determine the value of a patented component of a washing machine that shuts off the machine automatically in response to overflow (automatic shutoff). The respondent will be presented with a series of choices between washing machines that may, for example, vary on brand (LG, General Electric, Whirlpool), number of settings (eight, ten, twelve), color (white, beige, stainless steel), capacity (3, 4, 4.5 cu. ft.), and price ($400, $598, $648). Each respondent makes their choice in response to several sets (profiles) of products where the attributes (e.g., brand, number of settings) have randomly assigned levels (e.g., LG, General Electric, Whirlpool; eight, ten, twelve) for each profile. A regression is then used to compute part worths, the value that each feature adds to the total product. As in any survey, it is crucial to identify the appropriate universe of individuals likely to purchase the product. For instance, it would be inappropriate to conduct the washing machine conjoint survey on college students who can be assumed to have no intention of purchasing a washing machine in the near future.

Although conjoint analysis offers a potentially useful method for assessing WTP for a single feature of a multifeature product as a result of its experimental nature, the design of the profiles can lead to misleading results. The first issue is what features to include in the product profiles. If important features of the product are omitted (e.g., in the washing machine example, whether the machine is front-loading or top-loading) and the profiles include predominantly less important features e.g., door shape as round versus square, maximum spin speed), the estimated values for the component of interest are likely to be inflated.[166] Thus, an important preliminary step by the expert designing a conjoint analysis is to assess the importance of the various features of the product.[167]

164. Bernard Chao & Sydney Donovan, *Does Conjoint Analysis Reliably Value Patents?*, 58 Am. Bus. L.J. 225 (2021), https://doi.org/10.1111/ablj.12182. *See also* Suneal Bedi & David Reibstein, *Damaged Damages: Errors in Patent and False Advertising Litigation,* 73 Ala. L. Rev. 385 (2021); David Franklyn & Adam Kuhn, *The Problem of Mop Heads in the Era of Apps: Toward More Rigorous Standards of Value Apportionment in Contemporary Patent Law,* 98 J. Pat. & Trademark Off. Soc'y 182 (2016).

165. Some conjoint analyses ask respondents for rankings or ratings, but asking them to make choices is the most generally accepted approach because it is more reflective of what they would do in the marketplace. Bedi & Reibstein, *supra* note 164, at 399 nn.76 & 77.

166. Bedi & Reibstein, *supra* note 164.

167. In *Apple, Inc. v. Samsung Electronics Co.,* No. 11-CV-01846-LHK, 2014 U.S. Dist. LEXIS 29721 (N.D. Cal. Mar. 6, 2014), the conjoint survey used by Apple included seven attributes;

An additional challenge in designing a conjoint survey is to ensure that the features are described clearly and accurately, while simultaneously not drawing substantially more attention to features than consumers might naturally devote to them in the real world. This allows respondents to understand what each feature is and ensures that their understanding matches the meanings and definitions used in the case.[168] If a feature is unclear (e.g., electronic spinner), the respondent cannot make an informed decision that takes that feature into account. It also is essential, as with any survey, that the population be carefully chosen.[169]

Further, if the conjoint analysis is being used to identify prices, there is some controversy on how to capture supply-side aspects of the pricing that reflect the market. Given that market conditions may have changed due to the infringement, this can present a challenge.[170]

Samsung critiqued it for excluding other important attributes such as brand name and battery life. In *MacDougall v. American Honda Motor Co.,* No. 20–56060, 2021 U.S. App. LEXIS 37780 (9th Cir. Dec. 21, 2021), the court of appeals reversed after the district court had rejected the use of a conjoint survey for "the reduction of the amount of vehicle attributes in the final survey from thirty-three to four . . . The more limited a consumer's choice of vehicle features, the more artificially inflated the importance of the remaining features. . . ." MacDougall v. Am. Honda Motor Co., No. SACV 17–1079 JGB, 2020 U.S. Dist. LEXIS 166786, at *22 (C.D. Cal., Sept. 11, 2020). Including thirty-three attributes may be too many, but at issue here was the lack of justification for the four chosen (it was unclear how the pretest led to choosing those four). On appeal, the court found that the attributes chosen go to weight and not admissibility; however, the lack of a reasonable basis for attribute choice can seriously undermine the value of the analysis. There often is an inevitable tradeoff between overwhelming respondents with too many attributes and capturing those that are most important. The expert should explicitly provide reasons for the specific inclusion and exclusion of attributes.

168. For instance, in *Price*, the court rejected an analysis that used the term "Kerantindose Pro-Keratin + Silk" to isolate the impact of "Kerantindose" and "Pro-Keratin" since it includes "+ Silk." 2020 Dist. LEXIS 153255. In *Allegra v. Luxottica Retail North America*, 341 F.R.D. 373 (E.D.N.Y. Dec. 13, 2021), the court rejected the use of a conjoint survey that it said failed to properly describe the allegedly fraudulent omission in the advertising.

169. In *Cardenas v. Toyota Motor Corp.,* 418 F. Supp. 3d 1090 (S.D. Fla. 2019), the case involved a defect in a nonhybrid Toyota. The defendants argued that hybrid purchasers should not have been included. The crucial question was whether purchasers of hybrid cars would differ in their assessments from nonhybrid purchasers, which the court determined was a question of weight rather than admissibility. The analysis of subgroup effects is tricky because respondent subgroups may differ in the evaluations of comparison points and thus analyses needed to account for those distinctions; *see* Thomas Leeper et al., *Measuring Subgroup Preferences in Conjoint Experiments*, 28 Pol. Analysis, 207–21 (2020), https://doi.org/10.1017/pan.2019.30.

170. J. Gregory Sidak & Jeremy O. Skog, *Using Conjoint Analysis to Apportion Patent Damages,* 25 Fed. Cir. Bar J. 581 (2016). More generally, how to estimate supply side considerations is a matter of some debate. In *Cardenas v. Toyota Motor Corp.*, the court acknowledged differing opinions on whether the appropriate pricing information should emulate actual market prices and sales during the class period or the probable prices and sales in the situation where the relevant attribute took on the value that was under contention (the relevant attribute in this case was an odor emitted from a car's HVAC system). The court ruled that the approach to incorporating supply-side information does not go to admissibility, but could affect weight. *See also* Colangelo v. Champion

Some aspects of conjoint analysis are controversial. These include whether it is always necessary to give respondents the choice of declining to select any of the products offered in the profiles presented (i.e., "none of the above"). If the relevant population includes people who may prefer not to make the purchase, inclusion is warranted, but note there is a danger that respondents may choose the "no purchase" option to avoid assessing the products.[171] Similarly, there is some dispute about how many features of a product can be listed (e.g., can respondents process and respond to more than seven features?).

As with any survey, the expert should be prepared to explain the choices made in constructing the survey design. In light of the complexity of conjoint survey experimental designs and the controversial nature of some of the decisions the expert needs to make, a detailed explanation of those decisions is warranted to guide the court in evaluating whether the survey is admissible (or what weight to give it).

Benefits and Limitations Associated with the Mode of Data Collection Used in the Survey

Four primary methods have traditionally been used to collect survey data: (1) in-person interviews, (2) telephone interviews, (3) mail questionnaires, and (4) internet surveys. The choice of any data collection method for a survey should be justified by its strengths and weaknesses.

Common across in-person, telephone, and internet surveys is the use of computer-assisted techniques.[172] The interviewer conducting a computer-assisted interview (CAI), whether by telephone (CATI) or face-to-face (computer-assisted personal interviewing, or CAPI), follows the computer-generated script for the interview and enters the respondent's answers as the interview proceeds. A primary advantage of CATI and other CAI procedures is that skip patterns can be built into the program. If, for example, the respondent answers "yes" when asked whether she has ever been the victim of a burglary, the computer will generate further questions about the burglary; if she answers

Petfoods USA, Inc., No. 6:18-CV-122 (LEK/ML), 2022 U.S. Dist. LEXIS 60489, at *25 (N.D.N.Y. Mar. 31, 2022) ("[T]here is a legitimate difference of opinions, both among judges and experts, about the significance of supply side information in calculating loss of value.") (citing *In re* Fisher-Price Rock 'n' Play Sleeper Mktg., 567 F. Supp. 3d 406, 415 (W.D.N.Y. 2021)).

171. Daniel McFadden, *Stated Preference Methods and Their Applicability to Environmental Use and Non-use Valuations*, *in* Contingent Valuation of Environmental Goods: A Comprehensive Critique, 153–87 (Daniel McFadden & Kenneth Train eds., 2017). *See also* Moshe Ben-Akiva et al., *Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-Based Conjoint Analysis*, 10 Founds. & Trends in Econometrics 1 (2019), http://dx.doi.org/10.1561/0800000036.

172. Wright & Marsden, *supra* note 1, at 13–14.

no, the program will automatically skip the follow-up burglary questions. Interviewer errors in following the skip patterns are therefore avoided, making CAI procedures particularly valuable when the survey involves complex branching and skip patterns.[173] CAI procedures also can be used to control for order effects by having the program rotate the order in which the questions or choices are presented and facilitate the implementation of complex experiments with many conditions by randomly generating many potential factors at once.[174]

CAI procedures also include audio computer-assisted self-interviewing (ACASI) in which the respondent listens to recorded questions over the telephone or reads questions from a computer screen while listening to recorded versions of them through headphones. The respondent then answers verbally or directly on a keypad. ACASI procedures are particularly useful for collecting sensitive information (e.g., illegal drug use and other HIV risk behavior).[175] This is also useful in face-to-face interviews where the respondent can avoid having to reveal their (potentially sensitive) answer to the interviewer since they enter it themselves.

All CAI procedures require additional planning to take advantage of the potential for improvements in data quality. When a CAI protocol is used in a survey presented in litigation, the party offering the survey should supply for inspection the computer program that was used to generate the interviews. Moreover, CAI procedures do not eliminate the need for close monitoring of interviews to ensure that interviewers are accurately reading the questions in the interview protocol and accurately entering the respondent's answers (or that the respondent is paying attention and using the program correctly in entering their own answers).

In-Person Interviews

Although costly, in-person interviews generally are the preferred method of data collection, especially when visual materials must be shown to the respondent under controlled conditions. When the questions are complex and the interviewers are skilled, in-person interviewing provides the maximum opportunity to clarify or probe. Unlike a mail survey, in-person, telephone, and web-based surveys have the capability to implement complex skip sequences (in which the respondent's answer determines which question will be asked next) and the power to control the order in which the respondent answers the questions. In-person

173. Willem E. Saris, Computer-Assisted Interviewing 20, 27 (1991).

174. *See, e.g.*, Intel Corp. v. Advanced Micro Devices, Inc., 756 F. Supp. 1292, 1296–97 (N.D. Cal. 1991) (survey designed to test whether the term *386* as applied to a microprocessor was generic used a CATI protocol that tested reactions to five terms presented in rotated order).

175. *See, e.g.*, P.C. Cooley et al., *Automating Telephone Surveys: Using T-ACASI to Obtain Data on Sensitive Topics,* 16 Computs. in Hum. Behav. 1 (2000), https://perma.cc/L94R-AXKY.

interviewers also can directly verify who is completing the survey, a check that is unavailable in mail and web-based surveys. As described below in the "Selection and Training of the Interviewers" section, appropriate interviewer training, as well as monitoring of the implementation of interviewing, is necessary if these potential benefits are to be realized. Objections to the use of in-person interviews arise primarily from their high cost or, on occasion, from evidence of inept or biased interviewers. The latter concern is somewhat mitigated by computer assistance, described above.

Telephone Interviews

Telephone surveys offer a comparatively fast and lower-cost alternative to in-person surveys and can be particularly useful when the population is large and geographically dispersed. Telephone interviews (unless supplemented with mailed materials) should be used only when it is unnecessary to show the respondent any visual materials. Thus, an attorney may present the results of a telephone survey of jury-eligible citizens in a motion for a change of venue in order to provide evidence that community prejudice raises a reasonable suspicion of potential jury bias.[176] Similarly, potential confusion between a restaurant called McBagel's and the McDonald's fast-food chain was established in a telephone survey. Over objections from defendant McBagel's that the survey did not show respondents the defendant's print advertisements, the court found likelihood of confusion based on the survey, noting that "by soliciting audio responses [the telephone survey] was closely related to the radio advertising involved in the case."[177] In contrast, when words are not sufficient because, for example, the survey is assessing reactions to the trade dress or packaging of a product that is alleged to promote confusion, a telephone survey alone does not offer a suitable vehicle for questioning respondents.[178]

176. *See, e.g.*, State v. Baumruk, 85 S.W.3d 644 (Mo. 2002) (overturning the trial court's decision to ignore a survey that found about 70% of county residents remembered the shooting that led to the trial, and that of those who had heard about the shooting, 98% believed that the defendant was either definitely guilty or probably guilty); State v. Erickstad, 620 N.W.2d 136, 140 (N.D. 2000) (denying change-of-venue motion based on media coverage, concluding that "defendants [need to] submit qualified public opinion surveys, other opinion testimony, or any other evidence demonstrating community bias caused by the media coverage"). For a discussion of surveys used in motions for change of venue, see Neal Miller, *Facts, Expert Facts, and Statistics: Descriptive and Experimental Research Methods in Litigation*, *Part II*, 40 Rutgers L. Rev. 467, 470–74 (1988); National Jury Project, *Jurywork: Systematic Techniques* (2d ed. 2008).

177. McDonald's Corp. v. McBagel's, Inc., 649 F. Supp. 1268, 1278 (S.D.N.Y. 1986).

178. *See* Thompson Med. Co. v. Pfizer Inc., 753 F.2d 208 (2d Cir. 1985); Inc. Publ'g Corp. v. Manhattan Mag., Inc., 616 F. Supp. 370 (S.D.N.Y. 1985), *aff'd without op.*, 788 F.2d 3 (2d Cir. 1986).

In evaluating the sampling used in a telephone survey, the trier of fact should consider:

1. Whether (when prospective respondents are not business personnel) some form of random-digit dialing[179] was used instead of or to supplement telephone numbers obtained from telephone directories, because a high percentage of all residential telephone numbers in some areas may be unlisted;[180]
2. What types of phones were included—that is, landlines, cell phones, or both. Over ninety-six percent of American adults live in households with cell phones, leading to a shift in sampling procedures that sometimes rely only on cell phones.[181]
3. Whether the sampling procedures required the interviewer to sample within the household or business, instead of allowing the interviewer to administer the survey to any qualified individual who answered the telephone;[182] and
4. Whether interviewers were required to call back multiple times at several different times of the day and on different days to increase the likelihood of contacting individuals or businesses with different schedules and, as a result, to reduce nonresponse levels.[183]

Telephone surveys that do not include these procedures may not provide precise measures of the characteristics of a representative sample of respondents, but may be adequate for providing rough approximations. The vulnerability of the survey depends on the information being gathered. More elaborate procedures are advisable for achieving a representative sample of respondents if the survey instrument requests information that is likely to differ for individuals with listed

179. "Random-digit" dialing provides coverage of households with both listed and unlisted telephone numbers by generating numbers at random from the sampling frame of all possible telephone numbers. James M. Lepkowski, *Telephone Sampling Methods in the United States*, *in* Telephone Survey Methodology 81–91 (Robert M. Groves et al. eds., 1988).

180. Studies comparing listed and unlisted household characteristics show some important differences. *Id.* at 76.

181. Between 2% and 3% of the U.S. population has only landline access. Stephen J. Blumberg & Julian V. Luke, Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January–June 2020 (U.S. Dep't of Health & Hum. Servs., Feb. 2021), at 4, https://perma.cc/5TSM-V88J; *see also* Courtney Kennedy et al., *Implications of Moving Public Opinion Surveys to a Single-Frame Cell-Phone Random-Digit-Dial Design,* 82 Pub. Op. Q. 279 (2018), https://doi.org/10.1093/poq/nfy016 ("Analysis of more than 250 survey questions show[ed] that when landlines [were] excluded, estimates change[d] by less than one percentage point, on average.").

182. This is a consideration only if the survey is sampling individuals. If the survey is seeking information on the household, more than one individual may be able to answer questions on behalf of the household.

183. This applies equally to in-person interviews.

telephone numbers versus individuals with unlisted telephone numbers, individuals rarely at home versus those usually at home, or groups who are more versus less likely to rely on landlines or cell phones.

The report submitted by a survey expert who conducts a telephone survey should specify:

- The procedures that were used to identify potential respondents, including both the procedures used to select the telephone numbers that were called and the procedures used to identify the qualified individual to question;
- The number of telephone numbers for which no contact was made; and
- The number of contacted potential respondents who refused to participate in the survey (and how many follow-up attempts were made).[184]

Like computer-assisted personal interviewing (CAPI),[185] computer-assisted telephone interviewing (CATI) facilitates the administration and data entry of large-scale surveys.[186] A computer protocol may be used to generate and dial telephone numbers as well as to guide the interviewer.

Mail Questionnaires

In general, mail surveys tend to be substantially less costly than both in-person and telephone surveys.[187] Procedures that raise response rates for mail surveys include multiple mailings, highly personalized communications, prepaid return envelopes, incentives or gratuities, assurances of confidentiality, first-class outgoing postage, and follow-up reminders.[188]

A mail survey will not produce a high rate of return unless the recruitment process begins with an accurate and up-to-date list of names and addresses for

184. Additional disclosure and reporting features applicable to surveys in general are described in the section titled "Completeness and Accuracy of All Relevant Information in the Survey Report" below.

185. *See infra* text accompanying note 205.

186. *See* Tourangeau et al., *supra* note 88, at 289; Saris, *supra* note 173.

187. *See* Chase H. Harrison, *Mail Surveys and Paper Questionnaires*, *in* Handbook of Survey Research, *supra* note 1, at 498, 499.

188. *See, e.g.*, Richard J. Fox et al., *Mail Survey Response Rate: A Meta-Analysis of Selected Techniques for Inducing Response*, 52 Pub. Op. Q. 467, 482–84 (1988), https://doi.org/10.1086/269125; Kenneth D. Hopkins & Arlen R. Gullickson, *Response Rates in Survey Research: A Meta-Analysis of the Effects of Monetary Gratuities*, 61 J. Experimental Educ. 52, 54–57, 59 (1992), https://doi.org/10.1080/00220973.1992.9943849; Eleanor Singer et al., *Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature*, 59 Pub. Op. Q. 66, 71 (1995), https://doi.org/10.1086/269458; *see generally* Don A. Dillman et al., Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method (3d ed. 2009).

the target population. Even if the sampling frame is adequate, the sample may be unrepresentative if some individuals are more likely to respond than others (i.e., differential nonresponse). For example, if a survey targets a population that includes individuals with literacy problems, these individuals will tend to be underrepresented. Open-ended questions are generally of limited value on a mail survey because they depend entirely on the respondent to answer fully and do not provide the opportunity to probe or clarify unclear answers. Similarly, if eligibility to answer some questions depends on the respondent's answers to previous questions, such skip sequences may be difficult for some respondents to follow. Finally, because respondents complete mail surveys without supervision, survey personnel are unable to prevent respondents from discussing the questions and answers with others before completing the survey or to control the order in which respondents answer the questions. Although skilled design of questionnaire format, question order, and the appearance of the individual pages of a survey can minimize these problems, if it is crucial to have respondents answer questions in a particular order, a mail survey cannot be depended on to provide adequate data.

Internet Surveys

Over the past two decades there has been a massive increase in the use of internet surveys. One recent analysis found that the overwhelming majority of expert reports reviewed in the trademark and false advertising space were based on online surveys.[189] As of 2024, "an estimated" 96% of U.S. adults have access to the internet, including 99% of those between the ages of eighteen and forty-nine.[190] Although concerns remain about internet survey administration, and especially internet administration through online panels, many of those concerns can be mitigated through quality-control measures.

The benefits of internet administration are clear. The cost of conducting surveys online is a small fraction of the cost of hiring human interviewers to speak with potential respondents; the speed with which a researcher can find, screen, and survey hundreds of respondents online is far greater than in non-internet-based environments (e.g., via the mail); and there is less risk of interviewer bias and error when questions are presented by computer on a screen and respondents

189. Kugler & Henn, *supra* note 71, at 293–94.

190. Pew Rsch. Ctr., Internet, Broadband Fact Sheet (Nov. 2024), https://www.pewinternet.org/fact-sheet/internet-broadband. Of those age fifty to sixty-four, 98% had internet access. The sixty-five-plus group was lower at 90%. The same research shows that 79% of the U.S. population has broadband internet at home, while an additional 15% do not have broadband but do have internet access via smartphone. In 2011, only 79% of Americans had internet access. *Id.*

type their own responses.[191] Further, internet surveys may be particularly appropriate when the goal of the researcher is to simulate everyday commercial conduct that now frequently occurs online.

Computer administration also allows careful randomization of survey items, branching survey logic, and the display of complicated visual and auditory stimuli. In addition, the structure permits the survey to remind, or even require, the respondent to answer a question before the next question is presented. A further advantage of computer-administered surveys over interviewer-administered surveys is that they eliminate interviewer error because the computer presents the questions and the respondent records their own answers.

One major question for internet surveys is the adequacy of the sampling approach. The evaluation of this will depend on the type of internet survey involved, because the samples used in web-based surveys vary in fundamental ways. At one extreme is the list-based web survey. This web-based survey is sent to a closed set of potential respondents drawn from a list that consists of the email addresses of the target individuals (e.g., all employees at a company where each employee has a known email address). Here there is no reason not to use the tools of probability sampling, described above.

At the other extreme is the self-selected web pseudosurvey in which web users in general, or those who happen to visit a particular website, are all invited to express their views on a topic and they participate simply by volunteering. Participants are very likely to self-select on the basis of the nature of the topic, substantially distorting the results. These self-selected pseudosurveys resemble reader polls published in magazines and do not meet standard criteria for legitimate surveys admissible in court.[192] Occasionally, proponents of such polls tout the large number of respondents as evidence of the weight the results should be given, but the size of the sample cannot cure the likely participation bias in such voluntary polls.[193]

Between these two extremes is a large category of web-based survey approaches that researchers have developed to address concerns about sampling bias and nonresponse error. Many of these use the professionally managed nonprobability panels described above. These companies recruit diverse panels of respondents. Some respondents come from other well-traveled sites; others are

191. Reg Baker et al., *Research Synthesis: AAPOR Report on Online Panels*, 74 Pub. Op. Q. 711, 739 (2010), https://doi.org/10.1093/poq/nfq048 ("Overall, the research reported here generally suggests higher data quality for computer administration than for oral administration.").

192. *See, e.g.*, Merisant Co. v. McNeil Nutritionals, LLC, 242 F.R.D. 315 (E.D. Pa. 2007) (report on results from AOL "instant poll" excluded).

193. *See* Mick P. Couper, *Web Surveys: A Review of Issues and Approaches*, 64 Pub. Op. Q. 464, 480–81 (2000), https://doi.org/10.1086/318641 (a self-selected web survey conducted by the National Geographic Society through its website attracted 50,000 responses; a comparison of the Canadian respondents with data from the Canadian General Social Survey telephone survey conducted using random digit dialing showed marked differences on a variety of response measures).

pursued because their qualifications make them valuable for marketing studies. When a researcher contracts with a panel provider, the company can sift through its database to invite an appropriate mix of people to participate in a given survey. Those invited do not know the topic of the survey in advance (i.e., they are not opting in based on the topic). The final sample can be balanced to match desired demographics either through recruitment quotas or through response weighting.[194] The researcher using such data should obtain the procedures used by the vendor to ensure the quality of the data and make it available to the opposing party and the trier of fact. For example, what steps did the vendor take to minimize fraudulent respondents[195] and minimize duplicative respondents? Did the vendor themselves outsource the sampling?[196]

Generally, a researcher conducting an internet survey must take active steps to ensure an honest and attentive sample. On the most technical level, use of CAPTCHA[197] questions can weed bots out from the sample; panel providers can audit their panels to verify participant demographics, or at least check that self-reported demographics are consistent over time;[198] and survey platforms, which host the actual survey questions, can use cookies and other means to prevent multiple submissions from the same person or computer.[199] As with many other issues in the online space, there is constant evolution in the fraud-detection domain. A researcher should ensure that they, and their panel provider, are using current best practices to detect fraudulent responses and be prepared to explain how these responses work and their known impact.

Moving beyond the technology and into the survey itself, the researcher should take further steps to ensure an attentive sample. In the trademark space, a

194. *See, e.g.*, Philip Morris USA, Inc. v. Otamedia Ltd., No. 02 Civ 7575 (GEL) (KNF), 2005 U.S. Dist. LEXIS 1259, at *10–11 (S.D.N.Y. Jan. 28, 2005).

195. Andrew M. Bell & Thomas Gift, *Fraud in Online Surveys: Evidence from a Nonprobability, Subpopulation Sample*, 10 J. Experimental Pol. Sci. 148–53 (2023), https://doi.org/10.1017/XPS.2022.8.

196. Peter K. Enns & Jake Rothschild, *Do You Know Where Your Survey Data Come From?,* Medium, May 2, 2022, https://perma.cc/BD9T-SK25.

197. CAPTCHA is an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart." The respondent may be asked to decipher a set of distorted letters or complete a categorization task that would be difficult to automate.

198. Courtney Kennedy et al., Pew Rsch. Ctr., Evaluating Online Nonprobability Surveys 36 (May 2, 2016), https://perma.cc/VJ4L-BTJF:

> Most panels are double opt-in, meaning potential panelists first enter an email address and then respond to an email sent from the panel provider in order to confirm the email account. Depending on the vendor, other quality control features include IP address validation and digital fingerprinting, which guards against a single person having multiple accounts in a given panel.

199. "The most common technique for identifying duplicate respondents is digital fingerprinting. Specific applications of this technique vary, but they all involve the capture of technical information about a respondent's IP address, browser, software settings, and hardware configuration to construct a unique ID for that computer." Baker et al., *supra* note 191, at 756 (emphasis omitted).

few checks are particularly common.[200] One method of measuring attention in a traditional likelihood-of-confusion survey is to include a free-response question prompting participants to explain a prior answer. If the participant responds with gibberish or irrelevant responses, they are either not paying attention or not taking the survey seriously.[201] A similar check is possible in a Teflon survey testing whether a mark is generic.[202] There, a participant is asked whether each in a list of terms is a "brand name" or a "common name." A participant who responds that all terms are brand names or all are common names, that is, who gives a straight-line response, should be excluded. They either do not understand the task or are not trying to do it. For surveys aimed at purchasers of a particular product, it is common to ask which, of a list of products, they have bought in the past X months. A researcher can include nonexistent products on those lists, screening out people who implausibly claim to have purchased them.

The science behind these efforts to ensure honest responding is ever-evolving. A researcher should employ some of the above mechanisms in any online survey, and should be prepared to justify their choices. A court should not treat the methods described here as a comprehensive checklist, however. No particular check is independently necessary or sufficient, but use of appropriate checks is required to ensure quality when the respondent is completing the survey online.

Mixed-Format Surveys

A final approach to data collection does not depend on a single mode, but instead involves a mixed-mode approach. By combining modes, the survey design may increase the likelihood that all sampled members of the target population will be contacted. For example, a person without a landline may be reached by mail or email. Similarly, response rates may be increased if members of the target population are more likely to respond to one mode of contact versus another. For example, a person unwilling to be interviewed by phone may respond to a written or email contact. If a mixed-mode approach is used, the questions and structure of the questionnaires are likely to differ across modes, and the expert should be prepared to address the potential impact of mode on the answers obtained.[203]

200. These checks are described in greater length in Kugler & Henn, *supra* note 71, at 300–06.

201. Nat'l Fin. Partners Corp. v. Paycom Software, Inc., No. 14 C 7424, 2015 U.S. Dist. LEXIS 74700, at *25–26 (N.D. Ill. June 10, 2015) (questioning the results of a survey in part because the expert did not appear to have even read the free response answers, which included responses like "cool" and "LOL," when the expert should have excluded nonresponsive answers).

202. *See* E. Deborah Jay, *Genericness Surveys in Trademark Disputes: Under the Gavel*, *in* Trademark and Deceptive Advertising Surveys 107, 113–16, 120–25 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

203. Don A Dillman & Benjamin L. Messer, *Mixed-Mode Surveys*, *in* Handbook of Survey Research, *supra* note 1, at 550, 553.

Surveys Involving Interviewers

Selection and Training of the Interviewers

When interviews are used to question survey respondents, the results are trustworthy only if "sound interview procedures were followed by competent interviewers."[204] Properly trained interviewers receive detailed written instructions on everything they are to say to respondents, any stimulus materials they are to use in the survey, and how they are to complete the interview form. These instructions should be made available to the opposing party and to the trier of fact. Interviewers should be instructed to record verbatim the respondent's answers, to indicate explicitly whenever they repeat a question to the respondent, and to record any statements they make to the respondent or supplementary questions they ask.

Interviewers require training to ensure that they can follow directions in administering the survey questions and employ optimal interviewing techniques (e.g., pausing to give the respondent enough time to answer, resisting invitations to express their own beliefs or opinions, knowing when and how to use probes). Although procedures vary, there is evidence that interviewer performance suffers with less than a day of training in general interviewing skills and techniques for new interviewers.[205] Additional training is needed when the interviewer is responsible for last-stage sampling (i.e., selecting the particular respondents to be interviewed in an unbiased fashion). Further, in-person interviews also should be conducted in situations without distractions and where others cannot overhear the answers. Failure to follow these dictates was one ground used by a court to reject as inadmissible a survey that purported to demonstrate consumer confusion.[206]

Some compromises may be accepted when surveys must be conducted swiftly. In trademark and deceptive advertising cases, the plaintiff's usual request is for a preliminary injunction, because a delay means irreparable harm. Nonetheless, careful instruction and training of interviewers who administer the survey, as well as monitoring and validation to ensure quality control[207] and complete disclosure of the methods used for all of the procedures followed, are crucial

204. Toys "R" Us, Inc. v. Canarsie Kiddie Shop, Inc., 559 F. Supp. 1189, 1205 (E.D.N.Y. 1983). *See also* Wisconsin v. Indivior IncNo. 16–5073, 2020 U.S. Dist. LEXIS 219949, at *58 (E.D. Pa. Nov. 24, 2020) (praising a survey expert for "training interviewers carefully on interviewing techniques and the subject matter of the survey").

205. Fowler & Mangione, *supra* note 139, at 117; Nora Cate Schaeffer et al., *Interviewers and Interviewing*, *in* Handbook of Survey Research, *supra* note 1, at 437, 460.

206. *Toys "R" Us*, 559 F. Supp. at 1204 (some interviews apparently were conducted in a bowling alley; some interviewees waiting to be interviewed overheard the substance of the interview while they were waiting).

207. *See* section titled "Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias" below.

elements that, if compromised, seriously undermine the trustworthiness of any survey.

Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias

Three methods are used to ensure that the survey instrument was implemented in an unbiased fashion and according to instructions. The first, monitoring the interviews as they occur, is done most easily when telephone surveys are used, but can occur in the field via recordings (if respondents consent to it). Second, validation of interviews occurs when respondents in a sample are recontacted to ask whether the initial interviews took place and to determine whether the respondents were qualified to participate in the survey. The standard procedure for validation of in-person interviews is to telephone a random sample of about 10% to 15% of the respondents.[208] This validation procedure does not determine whether the initial interview as a whole was conducted properly, but it warns interviewers that their work is being checked and can detect gross failures in the administration of the survey. In computer-assisted interviews, further validation information can be obtained from the timings that can be automatically recorded when an interview occurs.

A third way to verify that the interviews were conducted properly is to examine the work done by each individual interviewer. By reviewing the interviews and individual responses recorded by each interviewer and comparing patterns of response across interviewers, researchers can identify any response patterns or inconsistencies that warrant further investigation. When interviewers conduct the survey, the identity of the interviewer (or a unique code) should be included in the data provided to the opposing party and trier of fact.

When a survey is conducted at the request of a party for litigation rather than in the normal course of business, a heightened standard for validation checks may be appropriate. Thus, independent validation of a random sample of interviews by a third party rather than by the field service that conducted the interviews increases the trustworthiness of the survey results.[209]

208. *See, e.g.*, Davis v. S. Bell Tel. & Tel. Co., No. 89-2839-CIV-NESBITT, 1994 U.S. Dist. LEXIS 13257, at *16 (S.D. Fla. Feb. 1, 1994); Nat'l Football League Props., Inc. v. N.J. Giants, Inc., 637 F. Supp. 507, 515 (D.N.J. 1986).

209. In *Rust Environment & Infrastructure, Inc. v. Teunissen*, 131 F.3d 1210, 1218 (7th Cir. 1997), the court criticized a survey in part because it "did not comport with accepted practice for independent validation of the results."

Accuracy of Data Entry

Analyzing the results of a survey requires that the data obtained on each sampled element be recorded and often coded before the results can be tabulated and processed. Procedures for data entry should include checks for completeness, checks for reliability and accuracy, and rules for resolving inconsistencies. Accurate data entry is maximized when responses are verified by duplicate entry and comparison, and when data-entry personnel are unaware of the purposes of the survey.

The need for these checks and rules to control mistakes in data entry is particularly great when data collected from survey respondents are combined with data on those same respondents obtained from institutional databases or records (e.g., a survey of jurors is combined with court data on their jury service).[210] In such cases, both data entries and matches need to be closely scrutinized.

Disclosure and Reporting

Early Disclosure About the Survey Methodology and Results

Objections to the definition of the relevant population, the method of selecting the sample, and the wording of questions generally are raised for the first time when the results of the survey are presented. By that time, it is often too late to correct methodological deficiencies that could have been addressed in the planning stages of the survey. The plaintiff in a trademark case[211] submitted a set of proposed survey questions to the trial judge, who ruled that the survey results would be admissible at trial while reserving the question of the weight the evidence would be given.[212] The Seventh Circuit called this approach a commendable procedure and suggested that it would have been even more desirable if the parties had "attempt[ed] in good faith to agree upon the questions to be in such a survey."[213]

210. *See generally* Jan Van den Broeck et al., *Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities*, 2 PLoS Med 2(10): e267, https://doi.org/10.1371/journal.pmed.0020267. *See also* Mary R. Rose & Marc A. Musick, *How Can You Tell if There Is a Crisis? Data and Measurement Challenges in Assessing Jury Representation*, 98 Chicago-Kent L. Rev. 35, 42 (2023).

211. Union Carbide Corp. v. Ever-Ready, Inc., 392 F. Supp. 280, 291 (N.D. Ill. 1975), *rev'd*, 531 F.2d 366 (7th Cir. 1976).

212. Before trial, the presiding judge was appointed to the court of appeals, so the case was tried by another district court judge.

213. *Union Carbide*, 531 F.2d at 386. On one occasion, the Seventh Circuit recommended filing a motion in limine, asking the district court to determine the admissibility of a survey based on an examination of the survey questions and the results of a preliminary survey before the party undertakes the expense of conducting the actual survey. Piper Aircraft Corp. v. Wag-Aero, Inc.,

The *Manual for Complex Litigation, Second*, recommended that parties be required, "before conducting any poll, to provide other parties with an outline of the proposed form and methodology, including the particular questions that will be asked, the introductory statements or instructions that will be given, and other controls to be used in the interrogation process."[214] The parties then were encouraged to attempt to resolve any methodological disagreements before the survey was conducted.[215] Although this passage in the second edition of the *Manual* has been cited with apparent approval,[216] the prior agreement that the *Manual* recommends has occurred rarely, and the *Manual for Complex Litigation, Fourth*, recommends, but does not advocate requiring, prior disclosure and discussion of survey plans.[217] As the *Manual* suggests, however, early disclosure can enable the parties to raise prompt objections that may permit corrective measures to be taken before a survey is completed.[218]

Rule 26 of the Federal Rules of Civil Procedure requires extensive disclosure of the basis of opinions offered by testifying experts. However, Rule 26 does not produce disclosure of all survey materials, because parties are not obligated to disclose information about nontestifying experts. Parties considering whether to commission or use a survey for litigation are not obligated to present a survey that produces unfavorable results. Prior disclosure of a proposed survey instrument places the party that ultimately would prefer not to present the survey in the position of presenting damaging results or leaving the impression that the results are not being presented because they were unfavorable. Anticipating such a situation, parties do not decide whether an expert will testify until after the results of the survey are available.

Nonetheless, courts can encourage early disclosure and discussion even if they do not lead to agreement between the parties. In *McNeilab, Inc. v. American Home Products Corp.*,[219] Judge William C. Conner encouraged the parties to submit their survey plans for court approval to ensure their evidentiary value; the plaintiff did

741 F.2d 925, 929 (7th Cir. 1984). On another occasion, the parties jointly developed a survey administered by a neutral third-party survey firm. Scott v. City of New York, 591 F. Supp. 2d 554, 560 (S.D.N.Y. 2008) (survey design, including multiple pretests, negotiated with the help of the magistrate judge).

214. Manual for Complex Litigation, Second § 21.484 (1985).

215. *See id*.

216. *See, e.g.*, Nat'l Football League Props., Inc. v. N.J. Giants, Inc., 637 F. Supp. 507, 514 n.3 (D.N.J. 1986).

217. MCL 4th, *supra* note 41, § 11.493 ("including the specific questions that will be asked, the introductory statements or instructions that will be given, and other controls to be used in the interrogation process").

218. *See id*.

219. 848 F.2d 34, 36 (2d Cir. 1988) (discussing with approval the actions of the district court). *See also* Hubbard v. Midland Credit Mgmt., No. 1:05-cv-0216, 2009 U.S. Dist. LEXIS 13938 (S.D. Ind. Feb. 23, 2009) (court responded to plaintiff's motions to approve survey methodology with a critique of the proposed methodology).

so and altered its research plan based on Judge Conner's recommendations. Parties can anticipate that changes consistent with a judicial suggestion are likely to increase the weight given to, or at least the prospects of admissibility of, the survey.[220]

Completeness and Accuracy of All Relevant Information in the Survey Report

The completeness of the survey report is one indicator of the trustworthiness of the survey and the professionalism of the expert who is presenting the results of the survey. The American Association for Public Opinion Research (AAPOR), a professional organization that brings together producers and users of survey data, offers standards for disclosure to be reported with any survey results.[221] The following list, which draws on the AAPOR guidelines, describes the elements that a survey report generally should include:

1. The purpose of the survey, its research sponsor, and those who conducted the research;
2. A definition of the target population;
3. A description of how the sample was recruited, including the method of selection (and how much of the population is covered); population quotas used, if any; respondent compensation, if any; the method (mode) of interview; the number of callbacks; respondent eligibility or screening criteria and method; and other pertinent information (e.g., the name of the vendor, if used);
4. A description of the results of sample implementation, including the number of
   a. potential respondents contacted,
   b. potential respondents not reached,
   c. noneligibles,
   d. refusals,
   e. incomplete interviews or terminations, and
   f. completed interviews and final sample size;
5. The dates of data collection;
6. The exact wording of the questions used, including a copy of the actual questionnaire (and screenshots of the questionnaire as viewed by

220. Larry C. Jones, *Developing and Using Survey Evidence in Trademark Litigation*, 19 Memphis St. U. L. Rev. 471, 481 (1989).

221. AAPOR also provides professional ethics guidelines as well as a transparency initiative where survey organizations can be certified as publicly disclosing their basic research methods and making them public, https://perma.cc/4GPB-7QKK.

respondents if administered electronically), interviewer instructions, and visual exhibits, with any cross-condition variations or randomizations marked;[222]
7. A description of any weighting or estimating procedures used;
8. A description of data-processing procedures (e.g., validity checks such as measures of the time respondents took to complete the survey, any data-imputation procedures, criteria for removing any responses);
9. Estimates of the sampling error, where appropriate (i.e., in probability samples);
10. Statistical tables clearly labeled and identified regarding the source of the data, including the number of raw cases forming the base for each table, row, or column;
11. Copies of interviewer instructions, validation results, and code books, including coding instructions for free-response data;[223] and
12. Acknowledgment of limitations to the survey.

As a general rule, any research should provide such information or explain why doing so is not possible. Additional information to include in the survey report may depend on the nature of sampling design. For example, reported response rates along with the exact time each interview occurred may assist in evaluating the likelihood that nonresponses biased the results. In a survey designed to assess the duration of employee pre-shift activities, workers were approached as they entered the workplace; records were not kept on refusal rates or the timing of participation in the study. Thus, it was impossible to rule out the plausible hypothesis that individuals who arrived early for their shift with more time to spend on pre-shift activities were more likely to participate in the study.[224]

222. The questionnaire itself can often reveal important sources of bias. *See* Marria v. Broaddus, 200 F. Supp. 2d 280, 289 (S.D.N.Y. 2002) (court excluded survey sent to prison administrators based on questionnaire that began, "We need your help. We are helping to defend the NYS Department of Correctional Service in a case that involves their policy on intercepting Five-Percenter literature. Your answers to the following questions will be helpful in preparing a defense.").

223. Failure to supply this information substantially impairs a court's ability to evaluate a survey. *In re* Prudential Ins. Co. of Am. Sales Pracs. Litig., 962 F. Supp. 450, 532 (D.N.J. 1997) (citing the first edition of this manual). *But see* Fla. Bar v. Went for It, Inc., 515 U.S. 618, 626–28 (1995), in which a majority of the Supreme Court relied on a summary of results prepared by the Florida Bar from a consumer survey purporting to show consumer objections to attorney solicitation by mail. In a strong dissent, Justice Kennedy, joined by three other justices, found the survey inadequate based on the document available to the court, pointing out that the summary included "no actual surveys, few indications of sample size or selection procedures, no explanations of methodology, and no discussion of excluded results . . . no description of the statistical universe or scientific framework that permits any productive use of the information the so-called Summary of Record contains." *Id.* at 640 (Kennedy, J., dissenting).

224. *See* Chavez v. IBP, Inc., No. CT-01-5093-RHW, 2004 U.S. Dist. LEXIS 28837 (E.D. Wash. Aug. 18, 2004).

Researchers should also ask, and disclose, whether respondents have previously participated in similar prior surveys. Many nonprobability samples come from vendors with online panels where respondents participate in multiple surveys over time. There also are vendors who have established analogous online panels with probability samples.[225] Some evidence suggests that repeat respondents from these panels differ from fresh sample draws.[226] Vendors often do not provide information on prior participation—something about which a researcher can inquire prior to using a given vendor. Even if not, researchers can, on their own, attempt to identify "professional respondents" and consider whether they might bias inferences (e.g., by measuring prior participation and evaluating its effect on outcomes).[227] If the expert asks participants whether they have participated in previous studies on the same topic, or obtains that information from other sources, the results obtained from the experienced and naive participants can be compared. Currently, many experts take the sensible approach of simply excluding participants who report having completed similar prior surveys. Regardless, researchers should always disclose prior participation and its possible effects or be explicit that such information does not exist and explain why it was not obtained.[228]

When the presenting party has access to the raw data from a survey, that data should generally be made available to opposing counsel upon request. This dataset should include responses from all participants whose data were ultimately used in the original expert's report and also all participants whose data were excluded by the original expert for reasons of quality after the completion of the survey, with an indication of why they were excluded (i.e., data-processing procedures).[229] If a researcher wishes to exclude unusually fast and slow responses, they should disclose this in the report and provide the data from those respondents as well.[230] This permits an opposing expert to rerun statistical analyses and determine if any debatable analytic choices made by the original expert had substantial effects on the report's conclusions.

225. Online Panel Research: A Data Quality Perspective (Mario Callegaro et al, eds., 2014).

226. Andrew Halpern-Manners & John Robert Warren, *Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey,* 49 Demography 1499 (2012), https://doi.org/10.1007/s13524-012-0124-x.

227. D. Sunshine Hillygus et al., *Professional Respondents in Nonprobability Online Panels*, *in* Online Panel Research: A Data Quality Perspective 219–37 (2014).

228. K.H. Jamieson et al., *Protecting the Integrity of Survey Research*, 2 PNAS Nexus 1 (2023), https://doi.org/10.1093/pnasnexus/pgad049. More generally, these authors supplement AAPOR's code by suggesting researchers should provide details on question order, details on respondent attrition (i.e., dropping out of the survey), and so on.

229. For example, an expert may exclude respondents who responded with gibberish to free-response items or who completed the survey much more quickly than did other participants.

230. The issue of completion time-based exclusions is explored at greater length in Kugler & Henn, *supra* note 71, at 305–06. At present, there is no agreed-upon standard for identifying "too fast" responses; most researchers appear to be using rules of thumb (for example, excluding those who took less than one third the median time). *Id.*

The public (and, presumably, also judges) sometimes express concern that surveys are not sufficiently reliable, generally citing polling accuracy in recent elections.[231] Election pollsters have the unenviable task of modeling a changing electorate's intended behavior in a circumstance where it matters a great deal whether the correct answer is 49% support or 51% support. Other surveys do not require this level of precision relative to a specified standard (i.e., greater than 50%). Rather, the relevant estimate the survey is seeking to reflect is whether approximately 20% of people (vs. 40%, 60%, or even 5%) in the target market are confused by a given advertisement. Thus, the accuracy of a survey estimate should be assessed in light of the degree of precision needed for the context.

Though transparency is very important, some information must be removed from data files before they are shared beyond the survey expert to protect participant confidentiality. All identifying information, such as the respondent's name, address, and telephone number, should be omitted. In the case of internet surveys, it is appropriate to also remove the participant's panel ID number, IP address, and precise geolocation information; these might all be linked to the participant's identity.[232] Keeping the survey duration and survey start time in the file may aid analysis, however, and will not compromise participant anonymity in most cases.

Greater efforts to promote participant anonymity are appropriate in cases where the population surveyed is small or otherwise susceptible to easy reidentification. If a survey is conducted of a workplace, for example, it may be that the total population numbers only in the hundreds. The identity of a respondent of any given age, ethnicity, and gender may therefore be narrowed down to one of only a few possibilities.[233] Much greater care should be taken to protect participant anonymity in such cases. Possibilities include omitting data fields that the opposing expert does not expect to use in their own analyses, or restricting access

231. *See, e.g.*, Scott Keeter et al., Pew Rsch. Ctr., *What 2020's Election Poll Errors Tell Us About the Accuracy of Issue Polling* (Mar. 2, 2021), https://perma.cc/LDF9-KUWK (commenting on this concern and evaluating it). For a discussion of the accuracy of 2020 polling, see American Association of Public Opinion Research, *Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls,* https://perma.cc/64K9-GVDJ (not reaching firm conclusions, but suggesting that this may have been due to greater nonresponse among Trump voters and difficulty in accounting for new voters in screening).

232. IP addresses are sometimes used to detect whether individuals took the survey more than once. If so, shared data should indicate which respondents shared an IP address but should generally not include the IP address itself.

233. This problem also arises in the case of privacy statutes such as HIPAA and FERPA, where deidentified data can be shared but identifiable data is tightly restricted. The Department of Education reviews a number of questions that may be relevant to a school or workplace survey in their FERPA guidance, https://perma.cc/CCD3-47Y8 (updated May 2013).

to the data to ensure that it is not shared beyond the opposing expert—and particularly not with the ultimate client.[234]

The respondents questioned in a survey generally do not testify in legal proceedings and are unavailable for cross-examination. Indeed, one of the advantages of a survey is that it avoids a repetitious and unrepresentative parade of witnesses. To verify that interviews occurred with qualified respondents, standard survey practice includes validation procedures,[235] the results of which should be included in the survey report.

Conflicts may arise when an opposing party asks for survey respondents' names and addresses so that they can re-interview some respondents. The party introducing the survey or the survey organization that conducted the research generally resists supplying such information.[236] Professional surveyors as a rule promise confidentiality to increase participation rates and encourage candid responses although, to the extent that identifying information is collected, such promises may not effectively prevent a lawful inquiry. Because failure to extend confidentiality may bias both the willingness of potential respondents to participate in a survey and their responses, the professional standards for survey researchers generally prohibit disclosure of respondents' identities. "The use of survey results in a legal proceeding does not relieve the Survey Research Organization of its ethical obligation to maintain in confidence all Respondent-identifiable information or lessen the importance of Respondent anonymity."[237] Although no surveyor–respondent privilege currently is recognized, the need for surveys and the availability of other means to examine and ensure their trustworthiness argue for deference to legitimate claims for confidentiality in order to avoid seriously compromising the ability of surveys to produce accurate information.[238] In

234. In the case of a workplace survey, the client might well be the current employer of the survey participant.

235. *See* section titled "Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias" above.

236. *See, e.g.*, Alpo Petfoods, Inc. v. Ralston Purina Co., 720 F. Supp. 194 (D.D.C. 1989), *aff'd in part and vacated in part*, 913 F.2d 958 (D.C. Cir. 1990).

237. CASRO, *supra* note 41, § I.A.3f; Am. Ass'n for Pub. Op. Rsch., AAPOR Code of Professional Ethics and Practices (revised Apr. 2021), https://perma.cc/DBC8-LTWJ.

238. United States v. Dentsply Int'l, Inc., No. 99–5 MMS, 2000 U.S. Dist. LEXIS 6994, at *23 (D. Del. May 10, 2000) (Fed. R. Civ. P. 26(a)(*1*) does not require party to produce the identities of individual survey respondents); *In re* Litton Indus., Inc., No. 9123, 1979 FTC LEXIS 311, at *13 & n.12 (June 19, 1979) (Order Concerning the Identification of Individual Survey-Respondents with Their Questionnaires) (citing Frederick H. Boness & John F. Cordes, *The Researcher–Subject Relationship: The Need for Protection and a Model Statute*, 62 Geo. L.J. 243, 253 (1973)); *see also* Applera Corp. v. MJ Rsch., Inc., 389 F. Supp. 2d 344, 350 (D. Conn. 2005) (denying access to names of survey respondents); Lampshire v. Procter & Gamble Co., 94 F.R.D. 58, 60 (N.D. Ga. 1982) (defendant denied access to personal identifying information about women involved in studies by the Centers for Disease Control based on Fed. R. Civ. P. 26(c) giving court the authority to enter

general, the better approach is for the opposing party to conduct their own survey rather than re-interrogate the participants in the previous one.

Concluding Observations

As this chapter has shown, judges have a variety of factors to consider when determining whether a survey should be admitted and how much weight it should be given. The *Manual for Complex Litigation, Fourth (MCL4)*, published in 2004, suggested a set of relevant factors for judges to evaluate.[239] In the past twenty years since the *MCL4* was published, survey methods have evolved (e.g., with the growth of internet and other computer technology, and recognition of the importance of survey-experimental methodology for causal inference), but these factors remain important. Thus, we close by presenting an updated list that clarifies and builds on the list of factors presented in the *MCL4*.[240]
Relevant factors include whether:

- the population was properly chosen and defined;
- the sampling frame and the sample itself were representative of that population;
- the data and details on data collection were fully and accurately reported;
- the survey questions asked were clear and not leading;
- the survey was conducted by qualified persons following proper procedures;
- the data were analyzed following accepted statistical principles;
- statements about causal relationships were supported by sufficient evidence about causality; and
- the results were presented in accordance with statistical standards.[241]

"any order which justice requires to protect a party or persons from annoyance, embarrassment, oppression, or undue burden or expense") (citation omitted).

239. MCL 4th, *supra* note 41, § 11.493.

240. The *Manual for Complex Litigation* distinguished between factors to be considered regarding admissibility and factors to be considered in assigning weight. As all of the factors can affect either admissibility or weight, depending on their quality, we have combined them in one set.

241. These include, where appropriate, information on sampling error and confidence intervals.

Glossary of Terms

The following terms and definitions were adapted from a variety of sources, including *Handbook of Survey Research* (Peter H. Rossi et al. eds., 1st ed. 1983; Peter V. Marsden & James D. Wright eds., 2d ed. 2010); *Measurement Errors in Surveys* (Paul P. Biemer et al. eds., 1991); Willem E. Saris, *Computer-Assisted Interviewing* (1991); Seymour Sudman, *Applied Sampling* (1976).

branching. A questionnaire structure that uses the answers to earlier questions to determine which set of additional questions should be asked (e.g., citizens who report having served as jurors on a criminal case are asked different questions about their experiences than citizens who report having served as jurors on a civil case).

CAI (computer-assisted interviewing). A method of conducting interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

CAPI (computer-assisted personal interviewing). A method of conducting face-to-face interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

CATI (computer-assisted telephone interviewing). A method of conducting telephone interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

closed-ended question. A question that provides the respondent with a list of choices and asks the respondent to choose from among them.

cluster sampling. A sampling technique allowing for the selection of sample elements in groups or clusters, rather than on an individual basis; it may significantly reduce field costs and may increase sampling error if elements in the same cluster are more similar to one another than are elements in different clusters.

confidence interval. An indication of the probable range of error associated with a sample value obtained from a probability sample.

conjoint survey. Survey-experiment designed to identify and estimate the causal effects of many treatment components simultaneously by using an orthogonal, fractional factorial design.

context effect. When a previous question influences the way the respondent perceives and answers a later question.

convenience sample. A sample of elements selected because they were readily available.

coverage error. Any inconsistencies between the sampling frame and the target population.

double-blind research. Research in which the respondent and the interviewer are not given information that will alert them to the anticipated or preferred pattern of response.

full-filter question. A question asked of respondents to screen out those who do not have an opinion on the issue under investigation before asking them the question proper.

mall intercept survey. A survey conducted in a mall or shopping center in which potential respondents are approached by a recruiter (intercepted) and invited to participate in the survey.

margin of error. An indication of the likely precision of an estimate from a probability sample; used to compute a confidence interval.

multistage sampling design. A sampling design in which sampling takes place in several stages, beginning with larger units (e.g., cities) and then proceeding with smaller units (e.g., households or individuals within these units).

nonprobability sample. Any sample that does not qualify as a probability sample.

open-ended question. A question that requires the respondent to formulate their own response.

order effect. A tendency of respondents to choose an item based in part on the order of response alternatives on the questionnaire (see primacy effect and recency effect).

parameter. See population value.

pilot test. A small field test replicating the field procedures planned for the full-scale survey; although the terms *pilot test* and *pretest* are sometimes used interchangeably, a pretest tests the questionnaire, whereas a pilot test generally tests proposed collection procedures as well.

population. The totality of elements (objects, individuals, or other social units) that have some common property of interest; the target population is the collection of elements that the researcher would like to study. Also, universe.

population value, population parameter. The actual value of some characteristic in the population (e.g., the average age); the population value is estimated by taking a random sample from the population and computing the corresponding sample value.

pretest. A small preliminary test of a survey questionnaire. See pilot test.

primacy effect. A tendency of respondents to choose early items from a list of choices; the opposite of a recency effect.

probability sample. A type of sample selected so that every element in the population has a known nonzero probability of being included in the sample; a simple random sample is a probability sample.

probe. A follow-up question that an interviewer asks to obtain a more complete answer from a respondent (e.g., "Anything else?" "What kind of medical problem do you mean?").

quasi-filter question. A question that offers a "don't know" or "no opinion" option to respondents as part of a set of response alternatives; used to screen out respondents who may not have an opinion on the issue under investigation.

random sample. See probability sample.

recency effect. A tendency of respondents to choose later items from a list of choices; the opposite of a primacy effect.

sample. A subset of a population or universe selected so as to yield information about the population as a whole.

sampling error. The estimated size of the difference between the result obtained from a sample study and the result that would be obtained by attempting a complete study of all units in the sampling frame from which the sample was selected in the same manner and with the same care.

sampling frame. The source or sources from which the objects, individuals, or other social units in a sample are drawn.

secondary meaning. A descriptive term that becomes protectable as a trademark if it signifies to the purchasing public that the product comes from a single producer or source.

simple random sample. The most basic type of probability sample; each unit in the population has an equal probability of being in the sample, and all possible samples of a given size are equally likely to be selected.

skip pattern, skip sequence. A sequence of questions in which some should not be asked (should be skipped) based on the respondent's answer to a previous question (e.g., if the respondent indicates that he does not own a car, he should not be asked what brand of car he owns).

stratified sampling. A sampling technique in which the researcher subdivides the population into mutually exclusive and exhaustive subpopulations, or strata; within these strata, separate samples are selected. Results can be combined to form overall population estimates or used to report separate within-stratum estimates.

survey-experiment. A survey with randomly assigned control and treatment groups, enabling the researcher to test a causal proposition.

survey population. See population.

trade dress. A distinctive and nonfunctional design of a package or product protected under state unfair competition law and the federal Lanham Act § 43(a), 15 U.S.C. § 1125(a) (1946) (amended 1992).

universe. See population.

References on Survey Research

Jean M. Converse & Stanley Presser, Survey Questions: Handcrafting the Standardized Questionnaire (1986).

Mick P. Couper, Designing Effective Web Surveys (2008).

Matthew DeBell, *Computation of Survey Weights*, *in* The Palgrave Handbook of Survey Research, 519 (David L. Vannette & Jon A. Krosnick eds., 2018).

Shari S. Diamond, *Control Foundations: Rationale and Approaches*, in Trademark and Deceptive Advertising Surveys: Law, Science and Design 239 (Shari S. Diamond and & Jerre Swann, eds., 2d ed. 2022).

Don A. Dillman, Jolene Smyth & Leah M. Christian, Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (3d ed. 2009).

James N. Druckman, Experimental Thinking: A Primer on Social Science Experiments (2022).

Experimental Methods in Survey Research: Techniques that Combine Random Sampling and Random Assignment (Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw & Brady T. West eds., 2019).

Arlene Fink, How to Conduct Surveys: A Step-By-Step Guide (4th ed. 2009).

Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer & Roger Tourangeau, Survey Methodology (2d ed. 2009).

Handbook of Survey Research (Peter V. Marsden & James D. Wright eds., 2d ed. 2010).

Matthew B. Kugler & R. Charles Henn, *Internet Surveys in Trademark Cases: Benefits, Challenges, and Solutions*, *in* Trademark and Deceptive Advertising Surveys: Law, Science and Design, 293 (Shari S. Diamond and Jerre B. Swann, eds., 2d ed. 2022).

Sharon Lohr, Sampling: Design and Analysis (2d ed. 2010).

Measurement Errors in Surveys (Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz & Seymour Sudman eds., 2004).

Online Panel Research: A Data Quality Perspective (Mario Callegaro, Reg Baker, Jelke Bethleham, Anja S. Göritz, Jon A. Krosnick & Paul J. Lavrakas eds., 2014).

The Palgrave Handbook of Survey Research (David L. Vannette & Jon A. Krosnick eds., 2018).

Questions About Questions: Inquiries into the Cognitive Bases of Surveys (Judith M. Tanur ed., 1992).

Howard Schuman & Stanley Presser, Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context (1981).

William Shadish, Thomas D. Cook & Donald T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inferences (2002).

Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur & Roger Tourangeau, Cognition and Survey Research (1999).

Seymour Sudman, Applied Sampling (1976).

Survey Nonresponse (Robert M. Groves, Don A. Dillman, John L. Eltinge & Roderick J. A. Little eds., 2002).

Telephone Survey Methodology (Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, James T. Massey & William L. Nicholls eds., 1988).

Roger Tourangeau, Lance J. Rips & Kenneth Rasinski, The Psychology of Survey Response (2000).

Trademark and Deceptive Advertising Surveys: Law, Science and Design (Shari S. Diamond & Jerre B. Swann eds., 2d ed. 2021).

Reference Guide on Estimation of Economic Damages

MARK A. ALLEN, CARLOS BRAIN, AND FILIPE LACERDA

Mark A. Allen, J.D., is Principal at Cornerstone Research.

Carlos Brain, M.B.A., Sc.M., is Vice President at Cornerstone Research.

Filipe Lacerda, Ph.D., M.B.A., Finance, is Vice President at Cornerstone Research.

The views expressed herein are solely those of the authors, who are responsible for the content, and do not necessarily represent the views of Cornerstone Research.

CONTENTS

FIGURE

TABLE

Introduction

This reference guide identifies areas of dispute that arise when economic losses are at issue in legal proceedings. We focus on explaining the issues in these disputes rather than taking positions on their proper resolution. We discuss the application of economic analysis within established legal frameworks for damages, cover topics in economics that arise in measuring damages, and, where useful, provide citations to cases that illustrate the principles and techniques discussed in the text.

We begin with a brief discussion of qualifications for damages experts. We then set forth the standard general approach to quantification of economic damages, with particular focus on translating the legal theory of the harmful event into an analysis of the economic impact of that event and choosing the measure of economic damages. Next, we consider disputes that may arise between the parties regarding damages estimation. Such disputes may concern basic issues (e.g., proper damages measure and losses caused), valuation issues (e.g., damages arising on single or multiple dates, and the use of market or other data to estimate damages), other issues (e.g., data validity, prejudgment interest, disaggregation), and limitations on damages. We also discuss damages in class actions as well as the application of damages principles to specific types of cases.

We use brief examples to illustrate the disputes that can arise and provide intuition for relevant economic issues. These examples are not full case descriptions; they are deliberately stylized, attempting to capture the types of disagreements about damages that arise in practical experience, although they are purely hypothetical. In many examples, the dispute involves factual and economic as well as legal issues. We do not try to resolve the disputes in these examples; hopefully the examples will help clarify the legal and factual disputes that need to be resolved before or at trial.

Damages Experts' Qualifications

Experts who quantify damages come from a variety of backgrounds. The expert should be trained and experienced in quantitative analysis. For economists, the common qualification is a Ph.D. degree. Damages experts with business or accounting backgrounds often have M.B.A. or other advanced degrees, or C.P.A. credentials. Both the method of damages estimation used and the substance of the damages claim dictate the specific areas of specialization the expert should have. In some cases, participation in original research and authorship of professional publications may add to the qualifications of an expert. However, relevant research and publications are not likely to be on the topic of damages measurement per se, but rather on topics and methods encountered in damages analysis.

Professional training and experience in areas relevant to the substance of the damages claim may be required. For example, in antitrust, a background in industrial organization (i.e., competition economics) may be helpful; in securities damages, a background in finance may assist the expert; and in the case of lost earnings, an expert may benefit from training in labor economics.

An analysis by even the most qualified expert may face a motion to exclude under Federal Rule of Evidence 702 as interpreted by the *Daubert* line of cases.[1] Under Rule 702, an expert witness who is "qualified . . . by knowledge, skill, experience, training, or education," is allowed to testify when

> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

These criteria, which are intended to exclude testimony that is irrelevant or is based on untested and unreliable theories, are often applied to damages analyses. Examples include the following:

- A proposed damages expert may be qualified generally to opine on damages issues but may lack the qualifications necessary to offer an opinion in the context of the particular matter at hand. For example, a labor economist may lack the proper education, training, and experience to opine on damages in a securities matter.
- A proposed expert may have used a methodology that is incorrect as a matter of economics; the results may have been based on insufficient data (e.g., too few observations); or a theoretically sound analysis may not have been properly applied, even if large amounts of data were used in the analysis.
- All or part of a damages analysis may no longer be relevant if certain legal claims have been dismissed, or expert opinion on damages simply may not be needed at all, such as where damages may be determined using a simple mathematical formula.

Even if a motion to exclude fails, it can be an effective way for a party to probe an opposing expert's damages analysis prior to trial. For this reason, among others, motions to exclude proposed expert damages testimony have become relatively commonplace.

1. Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993); Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999). For a discussion of emerging standards of scientific evidence, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

The Standard General Approach to Quantification of Economic Damages

In this section, we review the general principles of the standard approach to the estimation of economic damages and basic issues that arise in such estimation. For each principle, there are several areas of potential dispute. The sequence of issues discussed here is intended to identify most of the areas of disagreement between the damages analyses of opposing parties with respect to these principles.

General Principles

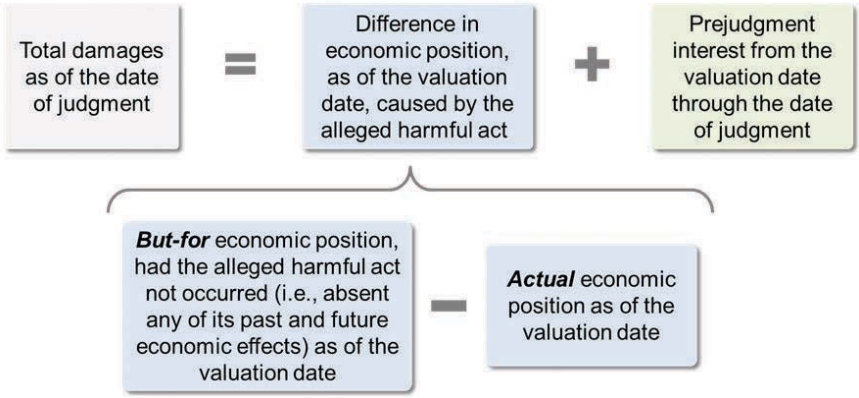
We begin by setting forth the standard general approach to damages quantification, with particular focus on defining the harmful event and the alternative, or but-for, scenario. In most cases, the goal of measurement of economic damages is to estimate the plaintiff's loss of economic value caused by the defendant's harmful act. The loss of value may arise from a single event, such as overpayment for a consumer product, or it may take the form of a reduced stream of profits or earnings over time. The losses are net of any costs avoided because of the harmful act.

In principle, then, economic damages are the difference between (1) the plaintiff's economic position assuming that the alleged harmful act had not occurred, and (2) the plaintiff's actual economic position as a result of the harmful act.[2] The former considers the plaintiff's economic position absent any past and future effects of the harmful act. We refer to this as the plaintiff's but-for economic position in the but-for world. The latter considers the plaintiff's actual economic position given the occurrence of the harmful act. We refer to this as the plaintiff's actual economic position in the actual world.

Figure 1 illustrates the structure of the standard damages study. We begin by estimating the plaintiff's losses as of the appropriate valuation date. The appropriate valuation date may vary depending on the nature of the loss and the plaintiff's theory of recovery (e.g., breach of contract, tort, fraud, etc.). Losses are the difference between the plaintiff's but-for economic position and the plaintiff's actual economic position as of the valuation date. Note that the plaintiff's but-for economic position may require estimating past losses as well as future losses, again

2. Damages are sometimes measured based on the defendant's gain rather than the plaintiff's losses. If measured based on the defendant's gain, economic damages are the difference between (1) the defendant's actual economic position as a result of the harmful act, and (2) the defendant's economic position assuming that the alleged harmful act had not occurred.

Figure 1. Damages studies generally adopt this framework. We therefore use it as the basic model for this reference guide.



determining their value as of the valuation date. The plaintiff may be entitled to prejudgment interest through the judgment date on their damages as of the valuation date.[3]

Isolating the Effect of the Harmful Act

The first step in a damages study is the translation of the legal theory of the harmful event into an analysis of the economic impact of that event. As noted above, in most cases, the analysis considers the difference between what the plaintiff's economic position would have been if the harmful event had not occurred and the plaintiff's actual economic position as of the valuation date.

The characterization of the harmful event begins with a clear statement of what occurred in the actual world. This characterization also will include a description of the but-for world, including the defendant's proper actions in place of their unlawful actions and a statement about the economic situation of all relevant parties absent the wrongdoing, with the defendant's proper actions replacing the unlawful ones. Damages measurement then determines the plaintiff's hypothetical economic position in the but-for world. Economic damages are the difference between that value and the actual value that the plaintiff achieved.

3. The plaintiff may also be due postjudgment interest, which is interest on the damages award from the date of judgment to the date the defendant satisfies the judgment.

Because the but-for world differs from what actually happened only with respect to the harmful act and its reasonably foreseeable economic consequences, damages measured in this way isolate the change in the plaintiff's economic position caused by the harmful act and exclude any change in the plaintiff's economic position arising from other causes.[4] Thus, a proper construction of the but-for world and measurement of the plaintiff's economic position in the but-for world by definition include in damages only the loss caused by the harmful act.

Properly translating the legal theory of the harmful act into an analysis of the economic impact of that event is a fundamental step in the estimation of economic damages. Failure to do so may result in the damages analysis being discounted or excluded altogether. For example, in *Comcast Corp. v. Behrend*, the U.S. Supreme Court found that the plaintiffs' expert had put forth a damages model that "failed to measure damages resulting from the particular antitrust injury on which . . . liability in this action is premised."[5] Citing the discussion in the prior edition of this reference guide, the Court determined that the expert's "methodology . . . identifies damages that are not the result of the wrong" and measures damages "caused by factors unrelated to an accepted theory of . . . harm."[6]

4. Reasonable foreseeability is a legal doctrine that courts use to determine whether the consequences of a legal or contractual breach are too remote to be borne by the defendant. Reasonable foreseeability is not a measure of damages, but rather an upper limit on damages. The use of "reasonable" here relates to the legal concept of the "reasonable person" or "economic person," and concerns actions that a reasonable person or economically rational actor under similar circumstances would take. *See, e.g.*, Richard A. Posner, Economic Analysis of Law (2d ed. 1977); Steven Shavell, Foundations of Economic Analysis of Law (2004); Israel Gilead & Michael D. Green, *Positive Externalities and the Economics of Proximate Cause*, 74 Wash. & Lee L. Rev. 1517 (2017); John Fabian Witt & Morgan Savige, *Foreseeability Conventions*, 44 Cardozo L. Rev. 1075 (2023); Russ VerSteeg, *Perspectives on Forseeability in the Law of Contracts and Torts: The Relationship Between "Intervening Causes" and "Impossibility,"* 2011 Mich. St. L. Rev. 1497 (2011). Reasonable foreseeability is not a term of art in the field of economics. While determinations regarding reasonable foreseeability may impact the damages analysis, application of the concept is not within the purview of the damages expert. As such, the damages expert may require assumptions from counsel as to whether a particular outcome is a reasonably foreseeable consequence of the alleged harmful act.

5. Comcast Corp. v. Behrend, 569 U.S. 27, 46 (2013).

6. *Id.* at 38 ("'The first step in a damages study is the translation of the *legal theory of the harmful event* into an analysis of the economic impact *of that event*.' The district court and the court of appeals ignored that first step entirely" (citations omitted)). *See also In re* POM Wonderful LLC, No. ML 10–02199 DDP RZX, 2014 WL 1225184, at *5 (C.D. Cal. March 24, 2014) (damages expert "made no attempt, let alone an attempt based upon sound methodology, to explain how Defendant's alleged misrepresentations caused any amount of damages"); *In re* Domestic Drywall Antitrust Litig., No. 13-MD-2437, 2017 WL 3700999, at *14 (E.D. Pa. Aug. 24, 2017) (expert's damages model was "riddled with assumptions that divorce the model from the facts").

General Categories of Damages Measures

In most cases, damages are measured based on one or more of the following six categories: (1) expectation, (2) reliance, (3) restitution, (4) statutory, (5) liquidated, and (6) punitive.[7] We discuss each of these in turn.

1. *Expectation*: Plaintiff restored to the same financial position as if the defendant had performed as promised.

Expectation damages typically apply to breach-of-contract claims, where the wrongdoing is the failure to perform as promised and the but-for scenario hypothesizes the absence of that wrongdoing—that is, proper performance by the defendant. Expectation damages are an amount sufficient to place the plaintiff in the same economic position as if the defendant had fulfilled the promise or bargain.[8]

2. *Reliance*: Plaintiff restored to the same position as if the relationship with the defendant or the defendant's misrepresentation (and resulting harm) had not existed in the first place.

Reliance damages generally apply to torts and to some breach-of-contract claims. Such damages restore the plaintiff to the same financial position they would have enjoyed absent the defendant's conduct.[9] Reliance most often includes out-of-pocket costs,[10] but may also include compensation for lost opportunities when appropriate.[11] In such cases, reliance damages may approach expectation

7. The law imposes several limitations on damages, which are discussed below in the section titled "Limitations on Damages."

8. *See* John R. Trentacosta, *Damages in Breach of Contract Cases*, 76 Mich. Bus. J. 1068, 1068 (1997) (describing expectation damages as damages that place the injured party in the same position as if the breaching party completely performed the contract); Spring Creek Expl. & Prod. v. Hess Bakken Inv., 887 F.3d 1003, 1026 (10th Cir. 2018) (defining expectation damages as the amount of damages necessary to place the injured party in the same position it would have occupied had the breach not occurred). *See also* Restatement (Second) of Contracts § 344(a) (1981).

9. *See* E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1148 (1979) (the objective of reliance damages is to put the promisee, or nonbreaching party, back into the position it would have been had the promise not been made). *See also* Restatement (Second) of Contracts § 344(b) (1981). Reliance damages include expenditures made in preparation for performance and performance itself. Restatement (Second) of Contracts § 349 (1981).

10. Out-of-pocket costs are expenses incurred by a party in preparing to perform (or in performing) a contract in reliance on the terms of the agreement. This is distinct from the notion of out-of-pocket damages used in securities fraud damages, which relates to the difference between price paid and a but-for price that would have been paid. *See* the discussion in the section titled "Estimation of Economic Damages in Securities Cases" below.

11. Compensation for lost opportunities embodies the economic concept of opportunity cost, which expresses the relationship between scarcity and choice. Given an economic actor facing a

damages.[12] For a tort, the goal of reliance damages is to place the plaintiff in a position economically equivalent to their position absent the harmful act.[13] For a breach of contract, measuring damages as the amount of compensation needed to return the plaintiff to the position they would have been in had the contract had not been made in the first place will result in refunding the part of the plaintiff's reliance investment that cannot be recovered in other ways. Thus, reliance damages may be appropriate where the plaintiff made an investment relying on the defendant's performance.

Example: Agent contracts with Owner for Agent to sell Owner's farm. The asking price is $1,000,000, and the agreed fee is 6%. Agent incurs costs of $1,000 in listing the property. A potential buyer offers the asking price, but Owner withdraws the listing. Agent calculates (expectation) damages as $60,000, the agreed fee for selling the property. Owner calculates (reliance) damages as $1,000, the amount that Agent spent to advertise the property.

3. *Restitution*: Plaintiff compensated by the amount of the defendant's gain from the unlawful conduct, also called compensation for unjust enrichment, disgorgement of ill-gotten gains, or compensation for unbargained-for benefits.

Restitution damages are awarded to prevent the defendant's unjust enrichment from the unlawful conduct. The goal of restitution damages is to restore the defendant to the economic position they would have been in had they not engaged in the wrongful act, and typically includes disgorgement of defendant's

choice between mutually exclusive alternatives, the opportunity cost of the alternative selected (e.g., entering into a contract with vendor A) is the value of the best alternative *not* chosen (e.g., entering into a contract with vendor B, C, or D). *See, e.g.*, N. Gregory Mankiw, Principles of Microeconomics 6 (8th ed. 2016) ("The opportunity cost of an item is what you give up to get that item." (emphasis omitted)).

12. In some situations, reliance damages can exceed expectation damages, but the court may limit damages to the expectation measure. *See, e.g.*, Fairholme Funds, Inc. v. Fed. Hous. Fin. Agency, No. 1:13-cv-1053-RCL, 2022 WL 11110584, at *4 (D.D.C. Oct. 19, 2022) (Plaintiff claimed reliance damages "many multiples higher" than the allowed measure of expectation damages. The court stated, "In contract cases, expectation damages are preferred, with reliance damages considered as an alternative or proxy if expectation damages are not readily ascertainable. . . . [P]laintiffs . . . cannot recover reliance damages so far in excess of ascertainable expectation damages that they would necessarily place those plaintiffs in a better position than they would have been in had the contract been performed.").

13. *See, e.g.*, East River Steamship Corp. v. Transamerica Delaval Inc., 476 U.S. 858, 873 n.9 (1986) ("tort damages generally compensate the plaintiff for loss and return him to the position he occupied before the injury").

ill-gotten gains.[14] In practice, restitution is often the same, from the perspective of quantification, as reliance damages. If the only loss to the plaintiff from the defendant's harmful act arises from an expenditure that the plaintiff made that cannot otherwise be recovered, the plaintiff receives compensation equal to the amount of that expenditure.[15]

4. *Statutory*: Plaintiff's compensation is an amount established by statute, either as a set amount per occurrence of wrongdoing or determined using a statutory formula. For example, this measure of damages is often found in cases involving violations of state labor codes and in copyright infringement.
5. *Liquidated*: Plaintiff compensated by an amount specified in a legal agreement or contract between the parties.
6. *Punitive*: Compensation to reward the plaintiff for detecting and prosecuting the wrongdoing or to deter similar wrongdoing in the future.

A plaintiff cannot normally seek punitive damages in a claim for breach of contract but may seek them in addition to compensatory damages in connection with a tort claim.[16] Although punitive damages are rarely the subject of expert testimony, economists have advanced the concept that punitive damages compensate a plaintiff who brings a case for a wrongdoing that is hard to detect or hard to prosecute.[17]

Basic Issues Arising in Damages Estimation

Losses Caused

The analysis of losses caused is typically a key part of the analysis of damages. Analyzing damages requires isolating the portion, if any, of the plaintiff's losses that can be attributed to the harmful act from the portion that can be attributed to other factors.

14. Note the distinction between restitution and reliance damages. The objective of restitution damages is to put the promisor, or breaching party, back in the position in which it would have been had the promise not been made. In contrast, reliance damages seek to put the promisee, or nonbreaching party, back in the position in which it would have been if the promise had not been made. *See* E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1148 (1979). Both measures seek to restore the status quo ante. *See also* Restatement (Third) of Restitution and Unjust Enrichment (2011).

15. *See* Restatement (Second) of Contracts § 344(c) (1981).

16. Compensatory damages are damages "sufficient in amount to indemnify [or compensate] the injured person for the loss suffered." *Damages*, Black's Law Dictionary (11th ed. 2019).

17. *See* Steven Shavell, Foundations of Economic Analysis of Law 244 (2004).

Example: Worker who smokes is the victim of a disease caused either by exposure to xerxium or by smoking. Worker makes leather jackets tanned with xerxium. The disease precludes Worker from earning wages. Worker sues the producer of the xerxium, Xerxium Mine, and calculates damages as all lost wages. Defendant Xerxium Mine, in contrast, attributes most of the losses to smoking and calculates damages as only a fraction of lost wages.

Frequently, the defendant will calculate damages on the premise that the harmful act had no causal relationship to the plaintiff's losses—that is, the plaintiff's losses would have occurred irrespective of the harmful act. The defendant's but-for scenario will thus describe a situation in which the losses happen anyway. This is equivalent to arguing that the harmful act occurred but the plaintiff suffered no incremental losses from it.

Example: Contractors conspired to rig bids in a construction deal. State seeks damages for subsequent higher prices. Contractors' damages estimate is zero because they assert that the only effect of the bid rigging was to determine the winner of the contract and that prices were not affected.

In other situations, unexpected events occurring after the harmful event (known as subsequent unexpected events) can increase or decrease the plaintiff's actual loss relative to what might have been expected at the time of the harmful event.

Example: Housepainter uses faulty paint, which begins to peel a month after the paint job. Owner measures damages as the cost of repainting. Painter disputes this measure on the ground that a hurricane that occurred three months after the paint job would have ruined a proper paint job anyway.

A plaintiff may argue that a harmful act has caused significant losses for many years. The defendant may respond that most of the losses that occurred from the injury are the result of causes other than the harmful act. Thus, the defendant may argue that the injury was caused by multiple factors, only one of which was the result of the harmful act, or that the plaintiff's injury was caused by subsequent events.

Example: Real Estate Agent is wrongfully denied affiliation with Broker. Agent's damages study projects past earnings into the future at the rate of growth of the previous three years. Broker's study projects that earnings would have declined even without the breach because of a downturn in the real estate market.

Comment: The difference between a damages study based on extrapolation from the past, here used by Agent, and a study based on actual data after the harmful act, here used by Broker, is one of the most common sources of disagreement in damages. This is a factual dispute that hinges on Broker demonstrating that there is a relationship between real estate market conditions and the earnings of agents. The example also illustrates how subsequent unexpected events can affect damages calculations.

Hypothesizing Legitimate Conduct of the Defendant in Analyzing the Plaintiff's Economic Position But For the Harmful Event

One party's damages analysis may hypothesize the absence of any act of the defendant that influenced the plaintiff, whereas the other party's damages analysis may hypothesize an alternative, legal act. This type of disagreement is particularly common in antitrust and intellectual property disputes. Although disagreement over the alternative scenario in a damages study is generally a legal question, opposing experts may have been given different legal guidance and therefore made different economic assumptions, resulting in major differences in their damages estimates.

Example: Defendant Copier Service's long-term contracts with customers are found to be unlawful because they create a barrier to entry that maintains Copier Service's monopoly power. Plaintiff Rival's damages study hypothesizes no contracts between Copier Service and its customers, so Rival would face no contractual barrier to bidding those customers away from Copier Service. Copier Service's damages study assumes medium-term contracts with its customers and argues that these would not have been found to be unlawful. Under Copier Service's assumption, Rival would have been much less successful in bidding away Copier Service's customers, and damages are correspondingly lower.

Comment: Assessment of damages will depend greatly on the substantive law governing the injury. The proper characterization of Copier Service's alternative conduct usually is an economic issue. However, the expert must also have legal guidance as to the proper legal framework for damages. Counsel for the plaintiff may instruct the plaintiff's damages expert to use a legal framework different from that used by counsel for the defendant.

Considering All the Differences in the Plaintiff's Situation in the But-For Scenario Versus Assuming That Many Aspects Would Be the Same as in Actuality

The analysis of some types of harmful events requires consideration of effects that involve changes in the economic environment caused by the harmful event, such as price erosion.[18] For a business, the main elements of the economic environment that may be affected by the harmful event include the prices charged by rivals, the demand facing the seller, and the prices of inputs. For example, misappropriation of intellectual property may cause lower prices because products produced with the misappropriated intellectual property compete with products sold by the owner of the intellectual property.[19]

In contrast, some types of harmful events do not change the plaintiff's economic environment. The theft of a small amount of the plaintiff's products would not change the market price of those products, nor would an injury to a worker change the general level of wages in the labor market. A damages study need not analyze changes in broader markets when the harmful act plainly has minuscule effects in those markets.

In situations where the plaintiff claims price erosion, the plaintiff may assert that absent the defendant's wrongdoing, a higher price could have been charged and therefore that the defendant's harmful act has eroded the market price. The defendant may reply that the higher price would lower the quantity sold. The parties may then dispute how much the quantity would fall as a result of higher prices.

Example: Valve Maker infringes Rival's patent. Rival calculates lost profits as the profits Rival would have made, including a price-erosion effect. The amount of price erosion is the difference between the higher price that Rival would have been able to charge absent

18. *See, e.g.*, Beijing Choice Elec. Tech. Co. v. Contec Med. Sys. USA Inc., No. 18 C 0825, 2020 WL 1701861, at ★9 (N.D. Ill. Apr. 8, 2020) ("To prove lost profits damages, a patentee must reconstruct the market that 'would have developed absent the infringing product,' and the alleged infringer's market share may be a relevant data point in this reconstruction. For instance, if an alleged infringer's market share increases after it introduces the product accused of infringement, that may shed light on what the patentee's market share would have been had there been no infringement. Or, if an alleged infringer's market share does not change significantly after it introduces the accused product, that may indicate that lost profits damages are not appropriate because consumers do not care whether the product uses the patented features at issue" (citations omitted)).

19. *See, e.g.*, Power Integrations, Inc. v. Fairchild Semiconductor Int'l, Inc., 711 F.3d 1348, 1382–83 (Fed. Cir. 2013) ("Lost revenue caused by a reduction in the market price of a patented good due to infringement is a legitimate element of compensatory damages. Indeed, an infringer's activities do more than divert sales to the infringer. They also depress the price [of the patented product].").

Valve Maker's presence in the market and the actual price. The price-erosion effect is that price difference multiplied by the combined sales volume of Valve Maker and Rival. Defendant Valve Maker counters that the volume would have been lower had the price been higher, and measures damages using the lower volume.

In more complicated situations, the damages analysis may need to focus on how an entire industry would be affected by the absence of the defendant's wrongdoing. For example, one federal appeals court held that a damages analysis for exclusionary conduct must consider that absent the defendant's conduct, firms other than the plaintiff may have entered the market. Competition from such firms would have reduced prices, and as a result the plaintiff's profits would have been lower than those posited in the plaintiff's damages analysis.[20]

Example: Printer Maker has used unlawful means to exclude rival suppliers of computer printer ink cartridges. Rival calculates damages on the assumption that they would have been the only additional seller in the market absent the exclusionary conduct and that Rival would have been able to sell its cartridges at the same price actually charged by Printer Maker. Printer Maker counters that other sellers would have entered the market and driven the price down, and so Rival has overstated their damages.

Comment: Increased competition lowers prices in all but the most unusual situations. Again, determination of the number of entrants that may have been attracted to the market absent the exclusionary conduct, and analysis of their effect on prices, requires a full economic analysis of the industry.

A clear statement of the plaintiff's situation but for the harmful event is helpful in avoiding double counting that can arise if a damages study confuses or combines reliance and expectation damages.

Example: Marketer is the victim of defective products made by Manufacturer; Marketer's business fails as a result. Marketer's damages

20. *See* Cave Consulting Grp., Inc. v. OptumInsight, Inc., No. 15-cv-03424-JCS, 2020 WL 127612, at *7 (N.D. Cal. Jan. 10, 2020) (no evidence provided that would provide the jury with any basis to disentangle "favorable aspects of an anticompetitive market such as an unnatural price differential between [plaintiff] and [its] competitors and limited competition from third parties because of the difficulty of entering the market."). *See also* Sun Microsystems Inc. v. Hynix Semiconductor Inc., 608 F. Supp. 2d 1166 (N.D. Cal. 2009) (holding that a before-and-after damages analysis requires "some showing that the market conditions in the two periods were similar but for the impact of the violation"); Dolphin Tours, Inc. v. Pacifico Creative Serv., Inc., 773 F.2d 1506, 1512 (9th Cir. 1985).

study adds together the out-of-pocket costs of creating the business and the projected profits of the business had there been no defects. Manufacturer's damages study measures the difference between the profits Marketer would have made absent the defects and the profits Marketer actually made.

Comment: Marketer has mistakenly added together damages from the reliance and expectation measures of damages. Under the reliance measure, Marketer is entitled to be put back to where they would have been had they not started the business in the first place. Damages would be total outlays less the revenues actually received. Under the expectation measure, Marketer is entitled to the extra profits they would have received had there been no product defects. Out-of-pocket expenses of starting the business would have no effect on expectation damages because they would be present in both the actual and but-for worlds and thus would offset each other.

Valuation Issues in the Analysis of Economic Damages

As discussed in the previous section, the analysis of economic damages compares a plaintiff's economic position in the but-for world—that is, absent any past and future effects of the alleged harmful act on the plaintiff's economic position—to that plaintiff's economic position in the actual world, and quantifies the economic value of that difference as of a specific date (the valuation date).[21] Because an analysis of economic damages is, by definition, a valuation exercise, it typically requires that economic experts resolve issues arising from the application of valuation approaches to specific case circumstances. This section introduces valuation issues often addressed by economic experts.[22]

21. As noted in the section titled "The Standard General Approach to Quantification of Economic Damages" above, in some situations the damages analysis aims to quantify the defendant's gain instead of the plaintiff's loss. In those situations, the relevant comparison is between the defendant's economic position in the actual world and the defendant's economic position in the but-for world—i.e., absent any past and future effects of the alleged harmful act on the defendant's economic position. The economic issues raised in this section similarly apply when analyzing the defendant's gain from the alleged harmful act.

22. In the section titled "Damages Issues in Selected Types of Cases" below, we introduce the analysis of economic damages in various case types, namely, consumer product liability, securities, antitrust, and loss of personal income. There, we do not specifically address business valuation disputes. However, the methodologies discussed in this section are also relevant to the analysis of economic damages in the context of business valuation disputes.

Issues Arising When the Potential Impact of the Alleged Harmful Act Is Analyzed Directly on a Particular Date

Issues Arising in Connection with the Use of the Market Price of the Product or Asset at Issue

Because a market price is simultaneously a price that the buying party was willing to pay at the time of the transaction and a price that the selling party was then willing to receive in exchange for the asset or product at issue, a market price provides an objective measure of economic value as of a particular point in time. However, as discussed below, the economic value represented by a market price reflects the specific circumstances under which the transaction occurred. Thus, there is often disagreement among experts as to whether and how the market price of the product or asset at issue can be used to analyze damages. This section summarizes some of the key issues that can generate disagreement.

The market price of the product or asset at issue in litigation is a frequent input to the analysis of damages on a particular date. A simple example of such a situation would be where the defendant's negligence caused total destruction of the plaintiff's cargo of wheat, worth $17 million at the then-current market price of wheat. In this simple example, a damages expert may calculate damages as just that amount, $17 million.

Market prices are often used to analyze the plaintiff's economic position in the but-for world, the actual world, or both. For example,

- In the simple example above, involving a destroyed cargo of wheat, the market price of wheat at the time of the defendant's negligence may be used to value the plaintiff's economic position but for the alleged harmful act (i.e., the $17 million value of the plaintiff's cargo of wheat had it not been destroyed).
- In a consumer fraud case involving a product falsely represented to have certain desirable characteristics, the market price of the product may be used to measure the plaintiff's actual economic position at the time of purchase (i.e., how much the plaintiff paid for the product).
- In a defamation case involving a publicly traded company plaintiff, the change in the company's stock price around the time of the defendant's defamatory act may serve as an initial input to the analysis of the difference in the plaintiff's economic position between the but-for world (e.g., using the stock price shortly before the defamatory act) and the actual world (e.g., using the stock price shortly after the defamatory act).

However, disagreement often arises among experts as to whether market prices (or changes in market prices over certain periods) serve as an adequate input to analyzing the impact of the alleged harmful act on the plaintiff's economic position.

First, when using market prices to value the plaintiff's economic position in the but-for world, disagreements may arise as to whether (and, if so, by how much) the market price that the expert relied on for their analysis represents a value that was itself impacted by the alleged harmful act. If the value represented by the market price reflects the impact of the alleged harmful act, such impact must be removed from the market price in order to properly value the plaintiff's economic position in the but-for world.

For example, if the destroyed cargo of wheat mentioned above was large enough that its destruction significantly constrained the supply of wheat, thereby increasing the market price of wheat, then the market price of wheat immediately following the destruction of the cargo of wheat will overstate the plaintiff's economic position but for the alleged harmful act and, thus, will overstate damages. In that case, disagreement may arise among the experts as to the extent to which the market price of wheat was affected by the defendant's negligence. Moreover, disagreement may arise with respect to what statistical or economic methodologies should be used to quantify the necessary adjustment to the market price in order to account for that impact.

Second, when market prices are used to analyze the plaintiff's economic position in the actual world, disputes may arise as to whether (and, if so, by how much) the market price fully reflects the impact of the alleged harmful act. If the market price does not fully reflect the impact of the alleged harmful act, then it needs to be adjusted using a reliable methodology to properly value the plaintiff's economic position in the actual world.

For example, suppose that a manufacturer of wood windows with publicly traded common stock treats its windows with a defective preservative, causing the windows to rot. The window manufacturer sues the preservative manufacturer for damages from lost sales and from the cost of replacing the defective windows. The window manufacturer's expert may be tempted to use the window manufacturer's stock price following the disclosure of the alleged harmful act as a measure of the manufacturer's economic position in the actual world. A potential problem with using the market price of the plaintiff's stock after the alleged harmful act is that the stock market may anticipate recovery in the form of a damages award, and this will attenuate at least some of the decline in price following the alleged harmful act.[23] In the extreme, if stock traders expect that

23. Note that this understatement of damages arises when the publicly traded company stands to recover a damages award as the plaintiff in a matter. However, changes in the market price of company stock have a different role in situations, such as a securities matter, where the public company is the defendant. Damages may be overstated by an unknown amount if the release of

the plaintiff will receive exactly full compensation, the plaintiff's market value is unlikely to change at all when knowledge of the wrongdoing—including the fact that a damages award will be made—hits the stock market. Thus, in this example, the use of the observed market price of the plaintiff company's stock at the time of the injury understates the actual amount of harm by an unknown amount, so the expert should consider using additional valuation techniques. Disputes may arise among the experts as to which valuation techniques should be used to quantify the harm and how they should be applied.

Third, when market prices are used to analyze the plaintiff's economic position, disputes may arise as to whether (and, if so, by how much) market prices reflect not just the impact of the alleged harmful act, but also the impact of other factors. Any impact of factors other than the alleged harmful act must be removed from the estimate of economic damages in order to isolate the harm caused by the alleged harmful act.

For example, consider a securities fraud matter where a mining company defendant is alleged to have overstated the size of its mineral reserves and where the company's stock price declined by 30% following its revelation of the true size of the mineral reserves. Suppose that, on the day the company revealed the true size of its reserves, it also announced that a natural disaster had caused its largest mine to collapse. To quantify the losses, if any, caused by the market learning of the company's true mineral reserves, it is necessary to disaggregate the impact of that disclosure from the impact of the disclosure of the mine collapse. A dispute may arise as to how the change in stock price should be adjusted to remove the impact of the disclosure of the mine collapse and isolate the impact of the disclosure of the company's true reserves.

Issues Arising in Connection with the Use of the Market Price of Ostensibly Similar Products or Assets

In many cases, experts cannot take the market price of the product or asset at issue and apply it directly to analyze the plaintiff's economic position in the but-for world or in the actual world. For example, there may not be a market for the product or asset at issue from which to obtain a market price (e.g., a company may not be publicly traded). However, experts may still be able to use the market prices of ostensibly similar products or assets—often labeled comparables—to analyze the economic value of the product or asset at issue.

previously fraudulently concealed adverse information causes a reduction in the value of the company both because of the adverse information and because the market anticipates that the company will pay a large damages award to investors who overpaid for their shares during the period when the information was concealed.

Identifying products or assets that are ostensibly similar to the allegedly affected product or asset

A common area of disagreement when relying on the market price of comparables to analyze damages is whether the identified comparables are economically similar to the allegedly affected product or asset. For example, in matters involving real estate, experts often identify comparables from the universe of nearby properties with similar characteristics (e.g., property size, condition, proximity to amenities, zoning restrictions). If the identified comparables are not similar to the product or asset at issue along economically relevant dimensions, the comparables' market prices may not serve as a reliable measure of the economic value of the allegedly affected product or asset (in the but-for world or in the actual world) and may lead the economic expert to significantly misestimate economic damages.

There may also be a dispute as to whether the period when the identified comparables were traded (and thus when their market price was determined) is too distant from the damages valuation date. Temporal distance between the damages valuation date and the period when the selected comparables were traded may cause the market prices of the selected comparables to reflect macroeconomic and business conditions that are significantly different from those in place as of the damages valuation date. In that case, the market prices of the comparable products or assets may not reliably measure the economic value of the product or asset at issue. For example, in a matter involving a sale of a private oil company that took place at a time when oil prices were at historic lows, evaluating the fairness of the sale price by comparing it to the prices at which similar companies were sold during an earlier period (when oil prices were significantly higher) could lead the expert to overstate the value of the company as of the sale date.

Adjusting the market price of ostensibly similar products or assets

Even when experts agree on a list of comparables to use in the damages analysis, disputes may arise as to whether (and, if so, how) the value of the allegedly affected product or asset implied by the comparables' market prices needs to be further adjusted to adequately measure the economic value of the product or asset at issue.

Accounting for differences between the ostensibly similar products or assets and the allegedly affected product or asset. Despite experts' best efforts, circumstances are often such that the best available comparables remain different from the product or asset at issue in economically significant ways. In such circumstances, additional adjustments to account for those differences may

be required to ensure that the economic value of the allegedly affected product or asset is not misestimated. For example, in a matter involving a business valuation, the business at issue may be much smaller than any of the identified comparable businesses, rendering value comparisons inappropriate. Therefore, an expert may propose to restate the value of comparable businesses using valuation ratios—a ratio of some measure of value, such as stock price, to some measure of economic output, such as earnings—and multiply the comparables' valuation ratios by the measure of economic output for the smaller business to estimate its value, thereby adjusting for size.

However, disputes often arise over whether any adjustments made to the value of the comparables appropriately capture all significant economic differences between the comparables and the product or asset at issue. In the business valuation example, a dispute may arise as to whether the adjustment based on valuation ratios fully accounts for differences between the company at issue and the comparable companies, such as differences in profitability, growth prospects, or riskiness.

Example: Oil Company deprives Gas Station Operator of the benefits of Operator's business. Oil Company's damages study starts by calculating the ratio of gas station sale value to gasoline sales for five nearby gas station businesses that have sold recently. The average ratio is $0.26 per gallon of sales per year. The Operator sells 1.6 million gallons per year, so the business was worth $0.26 × 1,600,000 = $416,000, according to the Oil Company's expert. The expert for the Operator argues that the sales used by the Oil Company occurred before a major business relocated nearby. Thus, the gas station sale value to gasoline sales should be increased to $0.30 to reflect the new growth rate as a result of the expected increase in business. The Operator's expert calculates the business to be worth $0.30 × 1,600,000 = $480,000.

Accounting for the impact of the harmful act. In some cases where experts rely on comparables to analyze damages, the alleged harmful act did not cause a total loss of the value of the product or asset at issue, but rather caused only a reduction in its value. In those cases, damages experts may adapt the valuation implied by the comparables to measure the loss caused by the alleged harmful act—that is, the difference between the plaintiff's economic position in the but-for world and in the actual world. For example, in a business valuation case, an expert may propose to use valuation ratios to analyze damages, as illustrated by the numerical example below, although disputes may arise as to their appropriateness.

Example: Oil Company breaches an earlier agreement with Gas Station Operator and opens another station near Operator's station. As a result, Operator's gasoline sales are reduced by 700,000 gallons per year. Oil Company's damages study applies an average valuation ratio (based on recent sales of gas station businesses) of $0.26 per gallon of sales per year to the reduction in sales: $0.26 × 700,000 = $182,000. In contrast, Operator's damages study uses a statistical analysis (also based on recent sales of gas station businesses) that finds the impact of sales on gas station value depends on the total sales volume of the gas station, such that, given the Operator's level of sales absent the breach, each additional (or marginal) gallon of *subtracted* sales reduces value by $0.47. As a result, the Operator's expert calculates damages of $0.47 × 700,000 = $329,000.

Comment: Because fixed costs associated with running a gas station (e.g., rent) are diluted by larger sales amounts, the average valuation of gasoline sales will be less than the marginal valuation. Thus, a damages model that accounts for the level of sales absent the breach is the conceptually correct approach.

Issues Arising in Connection with the Use of Models to Simulate the But-For Price

In certain cases, experts cannot adjust the market price of the asset or product at issue—or of ostensibly similar assets or products—to reflect the impact of the alleged harmful act. In these cases, some experts have used statistical and economic models to simulate market prices in the but-for world.

It is well established in economics that market prices are determined by supply and demand. To estimate market prices in the but-for world, experts have proposed models that seek to represent the price-formation process, or the outcome of that process, in the relevant market when supply and demand for the asset or product at issue exclude the effects of the alleged wrongdoing. Experts rely on assumptions that simplify some of the complexities of real markets and focus their analysis on the specific impact of the harmful act.

Experts also rely on a wide variety of data to estimate the relevant variables and parameters of their models and represent real markets to a reasonable degree of accuracy. These data can include market data such as prices and quantities sold of the asset or product at issue, or the prices and quantities sold of related assets or products (such as substitute and complementary products). Experts can also rely on nonmarket data, such as accounting data or data generated via surveys.

Disagreements occur with respect to the assumptions and data used in these models. This section summarizes certain issues that arise when experts rely on statistical and economic models to simulate or approximate but-for prices.

Realistic modeling of preferences, incentives, and constraints of buyers and sellers

Experts use economic and statistical models to simulate real markets or market outcomes by focusing on the analysis of certain key elements and relationships of the market and by making assumptions about the influence of other less relevant characteristics. For a model to be useful, it needs to provide a sufficiently realistic representation of the relevant market. This usually involves defining the key characteristics of supply and demand of the asset or product at issue. Disputes may arise over which assumptions about demand and supply more realistically reflect the relevant market in the but-for world.

A valid model must provide a sufficiently realistic representation of the preferences of buyers. Experts usually achieve this by relying on methods of demand estimation commonly used by economists and marketing academics. These methods are broadly classified in two categories. The first category, known as the revealed preference method, relies on data from the choices made and prices paid by buyers in the real world to infer the underlying preferences and restrictions of buyers. The second category, known as the stated preference method, relies on data from surveys of relevant subjects that are considered representative of the population of buyers of the asset or product at issue.

When data from surveys are used, disputes may arise about whether such data accurately represent respondents' preferences—respondents may behave differently when taking a survey compared to how they would behave in the real market. Disputes may arise about how realistically a survey represents the environment in which buyers make purchase decisions in the real world, or about how accurately the sample of survey respondents represents the relevant population of buyers. Refer to the *Reference Guide on Survey Research*, in this manual, for a detailed treatment of these and other issues that arise when surveys are conducted.

A common question that arises when experts rely on a model to represent demand for an asset or product is whether the model appropriately accounts for variation in the preferences of different groups or types of buyers. For example, in an intellectual property dispute, an expert could use an economic model to estimate the price of a brand-name drug covered by a patent in a but-for world where the patent was not enforceable and generic drug manufacturers were allowed to compete. An overly simplistic model of demand could lead to a prediction that the entry of generic drug manufacturers into the market would result in a decrease in the price of the brand-name drug. However, a more realistic demand model

that takes into account the preferences of different types of consumers could lead to a finding that the market price of the brand-name drug would increase as the smaller segment of consumers who are brand-conscious would be willing to pay more for the brand-name drug.

A valid model must also provide a sufficiently realistic representation of the preferences and incentives of sellers. In many cases, experts rely on simplifying assumptions to represent the characteristics of supply. Some experts have assumed that the supply quantity is fixed at the quantity that was sold in the actual world. Other experts rely on economic frameworks to account for the changes in the economic position of sellers. These frameworks may include simplifying assumptions about the general relationship between the quantity offered and the price offered in the market or may include sophisticated models that simulate the behavior of individual sellers. Disputes commonly arise about whether the assumptions made by an expert regarding supply reflect the true positions and incentives of the suppliers in the but-for world.

Defining a valid model of supply generally requires consideration of key characteristics of the market that may influence the prices that consumers pay. For example, in a dispute involving the alleged mislabeling of a prescription drug's potential side effects, an expert must consider that the prices that consumers pay for prescription drugs may be influenced by insurance companies that can negotiate prices with the manufacturer and that can adjust consumer prices based on copays and deductibles.

It may also be necessary to define the characteristics of key competitors, distributors, retailers, and other relevant market participants. For example, a model that attempts to measure the impact on ticket prices of an alleged anticompetitive agreement between two bus lines should take into account available alternative means of transportation, such as cars, trains, or possibly airplanes. Disputes may arise as to the number, type, and characteristics of relevant alternatives consumers may have.

Modeling the impact of the alleged misconduct on demand and supply

To be helpful in the estimation of damages, a model of but-for prices must adequately reflect supply and demand absent the alleged misconduct while excluding the influence of unrelated factors. In general, the more complex the allegations are, or the more complex the facts surrounding the allegations are, the more difficult it will be for an economic model to reliably estimate market conditions absent the alleged misconduct.

Complex allegations may require more flexible models to represent the but-for world adequately. In general, models that allow more flexibility and that

provide the expert with more control over how to represent demand and supply in the but-for world are more difficult to estimate, and require larger amounts of data. For example, consider a dispute between a large buyer of insurance and an insurance company in which the allegations involve whether the insurer properly invoked a complex set of conditions to cancel a portfolio of policies and avoid paying for them. The damages model would require significant flexibility to perform the valuation of the portfolio in a but-for world where the defendant insurer could invoke some, but not all, of the cancellation conditions that it did in the actual world. Disputes commonly arise about whether an expert model adequately reflects complex allegations.

A model used by an expert should be able to account for the impact of confounding factors unrelated to the allegations but that may result in the same type of harm alleged. For example, in a dispute about potential contamination of groundwater from an industrial plant, an expert could estimate a statistical model—known as a hedonic model—to predict home prices based on certain characteristics of a home (e.g., number of bedrooms, number of bathrooms, etc.). The expert could attempt to use sale price data of homes before and after the alleged contamination event to estimate the impact of the alleged contamination on home values. However, if a zoning change affecting the properties at issue was enacted around the same time as the alleged contamination event, the expert's model would need to distinguish any decrease in the price of homes from the alleged contamination from any decrease (or increase) due to the zoning change. Disputes commonly arise about whether an expert model adequately controls for the impact of confounding factors.

Surveys provide the expert with some control over how to represent the but-for world via the information conveyed to survey respondents. However, disputes may arise about the specific information that respondents are assumed to have in the but-for world. For example, an expert could propose a survey to analyze how consumers would react to a label disclosing the use of genetically modified corn on a package of cereal labeled "100% natural corn." Experts may disagree about the wording, size, and placement of such a label, all of which can affect the damages estimate arising from the survey.

Accounting for Market Frictions

When relying on market data to analyze economic damages as of a specific date, it may be necessary to consider how market frictions may have impacted market prices—whether it be the price of the allegedly affected product or asset or the price of ostensibly similar products or assets.

Perfectly competitive markets have what economists term a frictionless market structure. These markets have (1) a large number of buyers and sellers

of a single, homogeneous product; (2) fully informed participants; and (3) the feature that participants can easily enter or exit from the market. A friction is anything that prevents the market from being perfectly competitive. Economists have identified various types of market frictions that can impact market outcomes, including market power (e.g., oligopolies), adverse selection, and taxes.[24]

For example, markets for businesses and properties have frictions that may make transaction values depart from the usual concept of the price negotiated by a willing seller and a willing buyer. In the case of a forced sale, and thus a less willing seller, the transaction price may understate the value. Adverse selection, which occurs when one party knows more about a property or business than the other, may cause severe failures in some markets.[25] Sales prices tend to be lower when equipment with hidden defects cannot be distinguished from equipment in unusually good condition. This is because buyers are only willing to pay a lower price that reflects the possibility of hidden defects in the equipment and because owners of equipment in good condition may choose not to offer it at that lower price.

Example: Negligence of Tire Maker causes the total loss of a Boeing 737 airplane. Tire Maker's damages analysis uses the prices of 737s of similar age in the used airplane market to set a value of $23 million on the ruined airplane. However, Airline offers the testimony of an economist expert who explains that only a small fraction of 737s are ever put up for sale in the used airplane market. Rather, airlines choose to sell only defective airplanes because they continue to fly nondefective 737s. The expert then adjusts for the adverse selection of inferior airplanes in the used airplane market and places a value of $42 million on the airplane.

Comment: Airline expert's adjustment is merited in principle, although it is challenging to carry out and is likely to be the subject of expert disagreement.

In financial markets, differences in information available to different market participants or limited asset liquidity can lead to market prices that differ, sometimes significantly so, from the prices that would have prevailed in perfectly competitive financial markets. For example, the sale of a large block of shares of

24. A comprehensive treatment of economic frictions is outside the scope of this reference guide.

25. *See, e.g.*, George A. Akerlof, *The Market for "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q. J. Econ. 488 (1970), https://doi.org/10.2307/1879431.

a company's common stock may happen at a price that is much lower than recently prevailing market prices because potential purchasers of that block of shares are concerned that the seller might have negative, material nonpublic information about the value of the shares.

A major source of friction in property and business markets is the capital gains tax. Because capital gains are taxed only at realization, subtracting the amount of unrealized capital gains taxes from the value of a business or property would generally understate the value of the business or property to the existing owners if they have no plans to sell, except in the very distant future.

Issues Arising in Connection with the Use of Repair or Abatement Costs

In some cases, economic damages can be calculated as the costs that the plaintiffs incurred or would incur to restore their economic position from the actual world to the but-for world in which the alleged harmful act did not occur. These costs could include repair costs (e.g., the cost of the parts and labor used to repair a defective product), replacement costs (e.g., the cost of a third party hired to perform a task that the defendant was contracted to do but did not), or the costs of reducing or abating any harm caused by the alleged misconduct (e.g., the cleanup costs following an environmental contamination incident).

Damages experts often disagree on how to calculate these costs. First, experts may disagree on whether any costs allegedly incurred are attributable to the alleged misconduct. For example, if a service hired by the plaintiff to repair an allegedly defective product also performed general maintenance, the total cost paid may overestimate damages. Second, experts may disagree on the costs that are necessary to repair or abate the consequences of the alleged misconduct. For example, in a case where the plaintiff alleges that they will need to install expensive air filtration devices to reduce foul odors from a defective air conditioning system, experts may disagree on whether such costs are necessary if cheaper alternatives (e.g., more frequent maintenance) are available. Third, experts may disagree about whether the potential costs of repair match the economic value of the loss experienced by the plaintiff. For example, consider a builder who in the process of selling a house learned that a contractor used cheaper materials than promised. Assume that the cost of removing the cheaper materials and replacing them with the materials originally specified would be $100,000; however, the builder was able to sell the home by offering a discount of only $50,000. In this case, the replacement costs would overstate the economic loss suffered by the builder.

Issues Arising When the Potential Impact of the Alleged Harmful Act Spans Multiple Dates and Damages Are Analyzed Indirectly on a Single Date

Having discussed various issues that arise when experts analyze damages by assessing the impact of the alleged harmful act on a specific date directly, we now turn to issues arising in situations where economic experts analyze damages on a particular date indirectly. Specifically, we turn to situations where experts first assess the direct economic impact of the alleged harmful act (i.e., the flow of economic income or loss at a specific point in time that can be causally linked to the defendant's alleged misconduct) over multiple dates (e.g., the lost profits from operating a facility that was expected to produce widgets for a long period of time but was destroyed by a defendant's misconduct) and then value the cumulative direct impact over multiple dates as of a single valuation date.

Estimating the Period-by-Period Direct Impact of the Alleged Harmful Act

In situations where experts assess the direct economic impact of the alleged harmful act over multiple dates, it is necessary to estimate flows of economic income or loss to the plaintiff in past and future periods that resulted (or were expected to result) from the alleged harmful act. In doing so, experts face a number of considerations.

Many of those considerations, such as how the actions of plaintiffs and defendants would have differed had the alleged harmful act not occurred, or whether all economic consequences of the alleged harmful act have been considered, have been introduced more broadly in the section titled "Valuation Issues in the Analysis of Economic Damages" above. This section further explores these considerations in the specific context of assessing the period-by-period direct economic impact of the alleged harmful act.

Disputes about how the defendant's actions would have differed had the alleged harmful act not occurred

When analyzing damages, experts often rely on the plaintiff's allegations (e.g., as stated in a complaint or statement of claim) as the basis for identifying the alternative actions that the plaintiff alleges the defendant should have undertaken instead of the alleged harmful act. However, in some cases, the plaintiff's

allegations do not specify fully what the defendant could and should have done instead of the alleged harmful act, or the allegations are squarely contradicted by case facts. For example, consider a derivative lawsuit where a company's shareholders sue the CEO for using the firm's funds in wasteful company acquisitions that are expected to reduce the company's quarterly profits for periods lasting well into the future. The plaintiff's allegations may leave unspecified what the CEO should have done with the funds instead of the allegedly wasteful acquisitions or the allegations may assume that certain investment opportunities were available that the record demonstrates were not.

Depending on the circumstances, the ambiguity remaining in the allegations may be significant enough that it causes experts on opposite sides of the litigation to reach substantially different estimates of damages. For example, in the case of the CEO alleged to have engaged in wasteful corporate acquisitions, the expert retained by the CEO defendant may assume that the funds used to finance the acquisitions would have been used to fund internal projects that were expected to be as unprofitable as the wasteful acquisitions. On the other hand, the expert retained by the investor plaintiffs may assume that the funds would have been distributed to the shareholders as quarterly dividends instead. Thus, when assessing opposing experts' damages analyses, a critical issue may be how each expert modeled what the defendant's actions would have been had the alleged harmful act not occurred.

When the plaintiff's allegations are consistent with any of several alternative actions that the defendant might have taken had the alleged harmful act not occurred, there is a risk that expert analyses become tainted by hindsight. With the benefit of hindsight—often, after years have passed since the alleged harmful act occurred—the parties may instruct their experts to assume that the defendant would have taken particular actions (among various possible alternatives that are consistent with the plaintiff's allegations) that lead to a more favorable damages number. However, the defendant's actions, had the alleged harmful act not occurred, would have been based on the defendant's knowledge at the time of the alleged harmful act, and not on whatever knowledge only later became available, including as a result of litigation. In the example above involving wasteful corporate acquisitions, depending on what the CEO contemporaneously knew, had the alleged harmful act not occurred, the CEO's decision could have been either to pay out the funds as dividends (as the shareholders' expert assumed) or to reinvest the funds into internal projects (as the CEO's expert assumed).

Disputes about how the plaintiff's actions would have differed had the alleged harmful act not occurred

Even when a defendant's actions, had the alleged harmful act not occurred, are fully specified, there may be disagreements about what the plaintiff would have done in the hypothetical world where the alleged harmful act did not occur. Such

disagreements may also lead to significantly different damages numbers. Take the numerical example below, where the plaintiff in a real estate matter argued that undeveloped land was taken from him by the defendant's harmful act and that damages should include the value of the undeveloped land, as well as the value of the stream of monthly rental income for years to come that the plaintiff would have earned had the plaintiff been able to develop the land.

Example: Property Owner sues County for the value of undeveloped property condemned for a rapid-transit extension. Property Owner's damages claim is $18 million, which is the appraisal value of a hypothetical condominium development on the property less the anticipated cost of building the development. The County's expert, an appraiser, argues that the market value of the property is $2 million, based on comparable undeveloped land nearby.

Comment: In principle, if the real estate market is perfectly competitive, the current market value of undeveloped land and the market value of the same land with proper development, less the cost of that development, should be the same, because buyers would bid for the developed properties based on the value of the undeveloped land. Thus, Property Owner may have understated the development costs. But the value of nearby properties used as comparables may understate the value of the condemned property—they may be available for sale because they lack certain features that would make them more desirable to develop, such as a view. On the other hand, the Property Owner's valuation does not reflect the probability that the Property Owner may not succeed in building the condominium.

Moreover, as was the case with respect to modeling the defendant's actions had the alleged harmful act not occurred, there is a risk that expert analyses may be tainted by hindsight when making assumptions about what actions the plaintiff would have taken, had the alleged harmful act not occurred.

Disputes about whether all economic consequences of the alleged harmful act are being accounted for

Another area where disputes may arise when analyzing the direct economic impact of the alleged harmful act over multiple periods is whether the experts' damages analyses have fully considered all the reasonably foreseeable economic consequences of the alleged harmful act. Failure to fully consider all reasonably foreseeable economic consequences of the alleged harmful act may lead experts to estimate damages incorrectly, possibly significantly so.

For example, where the injury takes the form of lost sales volume, the plaintiff usually has avoided the cost of production for the lost sales. Thus, an expert would overstate damages by not including the avoided costs of production as a reduction in the estimated direct economic impact of the alleged harmful act. Calculation of these avoided costs is a common area of disagreement about damages. Conceptually, avoided cost is the difference between the cost that would have been incurred at the higher volume of sales but for the alleged harmful act and the cost actually incurred at the lower volume of sales achieved. The avoided-cost calculation is done for each period. The following are some of the issues that arise in calculating avoided cost:

- For a firm operating at capacity, expansion of sales is cheaper in the longer run than in the short run; however, if there is unused capacity, expansion of sales may be cheaper in the short run.
- The costs that can be avoided if sales fall abruptly are less in the short run than in the longer run.
- Avoided costs may include marketing, selling, and administrative costs as well as the cost of manufacturing.
- Some costs are fixed, at least in the shorter run, and are not avoided as a result of the reduced volume of sales caused by the harmful act.

Sometimes putting costs into just two categories is useful: those that vary with sales (variable costs) and those that do not vary with sales (fixed costs). This breakdown is approximate, however, and does not do justice to important aspects of avoided costs. In particular, costs that are fixed in the short run may be variable in the longer run. Disputes frequently arise over whether particular costs are fixed or variable. One side may argue that most costs are fixed and were not avoided by losing sales volume, whereas the other side may argue that many costs are variable.

Certain accounting concepts relate to the calculation of avoided cost. Profit-and-loss statements frequently report the cost of goods sold.[26] Costs in this category are frequently, but not uniformly, avoided when sales volume is lower. But costs in other categories, called operating costs or overhead costs, may also be avoided, especially in the longer run. One approach to the measurement of avoided cost is based on an examination of all of a firm's cost categories. The expert determines how much of each category of cost was avoided.

An alternative approach uses regression analysis or some other statistical method to determine how costs vary with sales as a general matter within the firm or across similar firms. The results of such an analysis can sometimes be used to measure the costs avoided by the decline in sales volume caused by the harmful act.

26. *See, e.g.*, United States v. Arnous, 122 F.3d 321, 323–24 (6th Cir. 1997) (finding that the district court erred when it relied on government's theory of loss because the theory ignored the cost of goods sold).

Disputes about the use of expected outcomes to estimate the impact of the alleged harmful act

In most situations where the expert's damages analysis includes an evaluation of the impact of the alleged harmful act on future flows of economic income or loss, it is necessary to account for the uncertainty that surrounds the actual amount of those flows under different, yet-to-be-determined economic circumstances. For example, in a matter involving the valuation of a business, it is typically necessary to analyze how the future profits of the business may differ under different states of the world (e.g., in an economic recession versus in an economic boom, or if the business's new product becomes popular versus if it does not).

A fundamental economic principle in the analysis of uncertainty is that the value of future flows of economic income or loss depends on the expected outcome of those flows.[27] The expected outcome of an uncertain future flow of economic income or loss is conceptually simple to calculate. First, it is necessary to determine the specific outcome for the economic flow under the various possible scenarios and the probability that each scenario will occur. Then, the expected outcome is calculated as the weighted average of the outcomes for the economic flow across the various possible scenarios, weighting each scenario by the probability that it will occur. For example, in a simple situation where a company faces a one-time future profit of $1 million with probability 40% and a profit of $16 million with probability 60%, the company's expected profit is $10 million ($1 million times 40% plus $16 million times 60%).[28]

Application of the expected-value approach involves a careful study of the various risk factors that are expected to impact the ultimate outcome for the flow of economic income or loss. For example, in a business valuation matter involving a commercial real estate developer, it may be necessary to consider how the company's profitability would differ depending on risk factors that are likely to include future sales prices of commercial real estate, construction costs

27. As discussed in the section titled "Disagreements About the Discount Rate to Value the Impact of the Alleged Harmful Act in Periods After the Valuation Date" below, another important issue in the analysis of damages when there are future flows of economic income or loss is the appropriate discount rate to be used to transform expected future flows into their present value.

28. Sometimes the alternatives and interactions between the possible outcomes become so complex that more sophisticated analytical methods may be required. In such cases, experts often generate hundreds to millions of possible outcomes using probabilistic simulation techniques (e.g., Monte Carlo). These techniques usually start from a model of the relation between the outcome variable of interest (e.g., a company's profits, or the payoffs from a financial derivative) and the risk factors that affect that variable (typically based on findings from the peer-reviewed academic literature or statistical analyses of historical data). Then, the value that the outcome variable of interest would take under different future realizations of the relevant risk factors is simulated. Because those realizations are not known with certainty, they are simulated based on random draws from the assumed or estimated statistical distribution of the risk factors. Those random draws are then used as an input to the model to determine the simulated value of the outcome variable of interest.

(e.g., labor, steel, or lumber), local zoning restrictions, and availability of bank financing. Disputes often arise among experts over which risk factors must be considered in analyzing expected values and with respect to how those factors are likely to impact the flows of economic income or loss under different situations. Economic experts often rely on peer-reviewed academic literature or on statistical analyses based on historical data to determine the relevant risk factors and how they are expected to impact the ultimate outcome.

Disputes about the economic situation in future periods

When it is necessary to estimate the direct period-by-period impact of the alleged harmful act in future periods, disputes often arise over the basis for the assumption that the future will look a certain way. For example, in matters involving the valuation of startup companies with no track record of earlier profitability, a common source of disagreement is the likelihood that the company at issue will reach profitability and, if it does, how profitable it will be. An expert who values a startup company with no track record of earlier profitability under the assumption that the company will reach profitability is likely to be scrutinized on the basis for that assumption.

Example: Plaintiff Xterm is a failed startup. Defendant VenFund has breached a venture-capital financing agreement. Xterm's damages study projects the profits it expected to make under its business plan. VenFund's damages estimate, which is much lower, is based on the value of the startup as revealed by sales of Xterm's equity made just before the breach.

Comment: Both sides confront factual issues in validating their damages estimates. Xterm needs to show that the expected profits according to its business plan were supported by relevant information available to it as of the time of the breach. VenFund needs to show that the sale of equity places a reasonable value on the firm (e.g., it did not take place in a "fire sale" and it incorporated the same information that was available as of the date of breach). This dispute can also be characterized as whether the plaintiff is entitled to expectation damages or reliance damages. Certain jurisdictions may limit damages for firms with no track record of profitability.[29]

29. *See, e.g.*, Schonfeld v. Hilliard, 218 F.3d 164, 172 (2d Cir. 2000) ("Although lost profits need not be proven with 'mathematical precision,' they must be 'capable of measurement based upon known reliable factors without undue speculation.' Therefore, evidence of lost profits from a new business venture receives greater scrutiny because there is no track record upon which to base an estimate. Projections of future profits based upon 'a multitude of assumptions' that require 'speculation and conjecture' and few known factors do not provide the requisite certainty." (citations omitted)).

Disputes about the consideration of non-cash effects of the alleged harmful act

In some situations, a careful assessment of the plaintiff's economic situation in the actual world and in the but-for world reveals that the direct economic impact of the alleged harmful act on a period-by-period basis may include the loss of a flow of economic income that does not directly involve the loss of actual cash.

For example, in some firms, employee stock options are a significant part of total compensation. Employee stock options grant their holders the right to purchase shares of company stock at a fixed price. Stock options are often used by early-stage businesses because the options do not require the business to pay out any cash at the time they are granted. At the same time, the options are valuable compensation for employees because they potentially allow them to benefit from the company's future growth. Specifically, at a future date, the options may be exercised, and the employee option holder will pay only the previously agreed exercise price (typically, the price per share at the time the options are received) as opposed to the price per share at the time the options are exercised. If the company continues to grow successfully into the future, such that the shares' market price vastly exceeds the exercise price, option holders will pay a price for new shares that is a fraction of the shares' market price.

In a lost-sales matter where a plaintiff company's employees are compensated using stock options, the parties may dispute whether the value of options should be included in the costs avoided by the plaintiff as a result of lost-sales volume. The defendant's expert might argue that stock options should be included, because their issuance is costly to the shareholders. The defendant might place a value on newly issued options and amortize this value over the period from issuance to vesting. The plaintiff, in contrast, might exclude options costs because the options cost the firm no cash payout at the time the options are granted. However, compensating employees using stock options imposes real economic costs on the firm's shareholders, in the form of actual value outlays when these options are exercised or in the form of expected future value outlays at the time the options are granted. Those costs should be considered.

Example: Firm A pays its sales manager $2,000 for every machine sold at $100,000, as well as options to purchase 1,600 shares in a year at the existing price of $10 per share. Firm A asserts that it lost $10 million in sales (100 machines) as a result of Firm B's disparagement of Firm A. In its damages analysis, Plaintiff Firm A states that the lost sales represent lost profits of $5.8 million: $10 million less $4 million in avoided production costs and $200,000 in avoided sales commissions. Defendant Firm B calculates that each stock option is worth $5 today based on an analysis

using accepted financial models to value the options. Thus, Firm B asserts that damages are $5 million: $10 million less $4 million in avoided production costs and $1 million in sales commissions ($200,000 plus $5 × 100 × 1,600).

Valuing the Period-by-Period Impact of the Alleged Harmful Act as of a Single Valuation Date

This section discusses issues that arise when determining the appropriate rates for the purposes of discounting and capitalization, as well as the appropriate valuation date. Once an expert generates estimates of the direct economic impact of the alleged harmful act in each relevant period, it is then necessary to convert those estimates into a single measure of damages as of a specific date—the damages valuation date. As noted above, experts do so by applying the fundamental economic principle that the value of flows of economic income or loss at different points in time can be translated into an equivalent value as of a particular date through discounting or capitalization using appropriate rates.

Discounting is the act of translating a future economic value to its equivalent economic value as of an earlier point in time. For example, discounting at an appropriate rate can be used to translate receiving $105 a year from now into its economic value today and to compare that to receiving $100 today. If the appropriate discount rate is 10%, then today's value of receiving $105 in a year is approximately $95—calculated as $105 divided by (1 + 10%)—and receiving $100 today is worth more than receiving $105 in a year. Capitalization is the reverse of discounting—that is, it is the act of translating a past economic value into its equivalent value as of a later point in time. For example, capitalizing at an appropriate rate can be used to compare receiving $100 today to having received $95 a year ago (and having reinvested that amount). If the appropriate capitalization rate is 10%, then today's value of $95 received a year ago is approximately $105—calculated as $95 multiplied by (1 + 10%)—and $95 received a year ago is worth more than $100 received today.

Disagreements about the discount rate to value the impact of the alleged harmful act in periods after the valuation date

As a matter of economics, the appropriate discount rate to convert the economic value of future flows of economic income or loss (e.g., $100 of business profits in one year) into their economic value as of an earlier date depends on two factors: the time value of money and the risk premium. The time value of money relates to how individuals and companies trade off sure dollars in the future for sure

dollars as of an earlier date, and it is typically measured using a risk-free rate. The risk premium relates to how individuals and companies trade off uncertain dollars for certain dollars. As the name suggests, the risk premium is typically expressed as an increment to the time value of money, such that the discount rate can be calculated as the sum of the time value of money (i.e., the risk-free rate) and the risk premium.

Both factors are important in the determination of discount rates and can have a substantial impact on damages estimates. This section discusses key issues faced by experts when estimating discount rates that adequately measure the time value of money and the risk premium.

Estimating the time value of money

As noted above, the time value of money relates to how individuals and companies trade off dollars in the future for dollars as of an earlier date. The value of money has a temporal dimension for various reasons. First, inflation erodes the value of future money by reducing the basket of goods and services that can be purchased with the same amount of dollars. Second, individuals and companies can reinvest today's dollars in order to obtain a larger number of dollars in the future. Thus, the time value of money on a given date depends on inflation expectations and on available investment opportunities. As those change over time, so does the time value of money. Discount rates used in damages analyses that do not adequately account for the time value of money as of the valuation date may generate unreliable damages estimates.

To estimate the time value of money, experts typically rely on interest rates derived from the prices of safe government debt (in whatever currency damages are being sought and measured). When damages are being measured in U.S. dollars, experts typically use interest rates derived from the prices of U.S. Treasury bills, notes, and bonds of various maturities (ranging from a few days to as many as thirty years). U.S. government debt is typically regarded as risk-free, in the sense that the investor is guaranteed the promised return on their investment if the security is held to its maturity.[30]

Interest rates derived from longer-term debt tend to differ from interest rates derived from shorter-term debt: Long-term interest rates are typically higher than short-term interest rates in normal economic conditions. Thus, the relevant interest rate to account for the time value of money depends on the horizon of the flows being analyzed. For example, lost profits thirty years into the future should be valued using discount rates that measure the time value of money based on interest rates with a similar horizon.

30. In truth, U.S. government debt is not entirely risk-free, because there is a risk (remote as it may be) that the U.S. government may default.

Estimating the risk premium

When determining the appropriate discount rate to value a flow of economic income or loss, the key source of disagreement among experts tends to be the estimation of the risk premium. Disagreements over this topic also tend to be difficult for the layperson to assess because the academic literature that studies risk premiums is vast and tends to be highly technical. Therefore, particular attention needs to be given to the basic principles surrounding the estimation of risk premiums in order to evaluate experts' positions and their bases.

The risk premium associated with a particular flow of economic income or loss is the compensation investors require for bearing the risks associated with that flow. The risks associated with a particular flow of economic income or loss are typically split into two types: systematic risks and idiosyncratic risks. Systematic risks relate to broader market conditions, that is, risks arising from factors that impact both the flow of economic income or loss at issue and the economy more broadly. Systematic risks affecting a firm include economic downturns, inflation, and other risks that an investor cannot avoid by holding a diversified investment portfolio. In contrast, idiosyncratic risks are risks arising from factors that impact the flow of economic income or loss at issue but do not impact the economy more broadly. Idiosyncratic risks may include whether a venture will succeed, a firm's ability to obtain financing, whether a competitor will develop a similar product, and risks related to the pricing of the product or competitive products, such as the price of inputs.

A basic principle of financial economics is that in order to make risky investments, investors require compensation for risks associated with those investments. Thus, whenever the uncertainty surrounding a flow of economic income or loss is related to overall market conditions, it is necessary to account for that uncertainty in the estimation of the discount rate.

In ideal competitive financial markets, investors demand compensation only for systematic risks, because those are the risks that investors cannot eliminate by holding a well-diversified portfolio of assets. Thus, experts often assume that no risk premium is associated with idiosyncratic risks.

A common approach for determining the risk premium associated with a particular flow of economic income or loss, although not the appropriate one under all circumstances where the calculation of a risk premium is warranted, is part of the Capital Asset Pricing Model (CAPM). The CAPM is a standard model in financial economics to analyze the relation between the risk from investing in a particular asset and the expected rate of return that investors demand for that asset (i.e., its discount rate).[31] Under the CAPM, the expected return for a given asset

31. *See, e.g.*, Aswath Damodaran, Damodaran on Valuation: Security Analysis for Investment and Corporate Finance 31–35 (2d ed. 2006); Richard A. Brealey, Stewart C. Myers & Franklin Allen, Principles of Corporate Finance 205–08 (13th ed. 2020).

($E[R_i]$) is equal to the risk-free rate (R_f) (which compensates investors for the time value of money) plus a risk premium for that specific asset (which compensates investors for taking on the risk from investing in the asset). The risk premium for the asset is equal to the risk premium for the overall market ($E[R_m] - R_f$) multiplied by the asset's *beta*, which is explained below. Mathematically, the CAPM equation for the expected return (or discount rate) on an asset is

$$E[R_i] = R_f + \beta_i \times (E[R_m] - R_f)$$

A key part of applying CAPM to the valuation of a particular asset is to determine the asset's beta (β_i). Beta measures the sensitivity of the asset's returns to changes in returns on the overall market. Assets whose returns are more sensitive to the returns of the overall market tend to decline more when the overall market declines. Because investors require compensation for such declines, assets with higher betas carry higher risk premiums (and, thus, have higher discount rates) than assets with lower betas. To illustrate, for a company whose stock has a beta of 1.5, if the index of value for a representative and broad set of firms[32] decreases by 10% over a year, then the value of the company's stock tends to decrease by 15% over that year (1.5 times 10%). Under CAPM, the risk premium for that stock can be calculated as the 1.5 beta multiplied by the historical average risk premium for the overall stock market.[33] The calculation may be presented in the following format:

1. Beta: $\beta_i = 1.5$
2. Market risk premium: $E[R_m] - R_f = 6.0\%$
3. Risk premium (line 1 times line 2): 9.0%

In this example, if the expert estimates the risk-free rate as 2.00%, for example, based on 20-year treasury rates as of the valuation date, then the estimated discount rate will be 11.00% (2.00% plus 9.00%).

Disagreements about the capitalization rate to value the impact of the alleged harmful act in periods before the valuation date

When the direct economic impact of the alleged harmful act is a flow of economic income or loss that occurred (or would have occurred in the but-for world) before the valuation date, it is necessary to determine a capitalization rate that can be used to value the past flow of economic income or loss as of the later valuation date. Such determination often requires careful consideration of the

32. *E.g.*, the S&P 500 index.

33. Experts may disagree on the appropriate methodology to estimate the market risk premium.

specific circumstances faced by the plaintiff that was impacted by the alleged harmful act at the time the direct economic impact occurred.

In some situations, it is natural to see past flows as income that could have been reinvested, and thus, it is necessary to consider the rate of return that investment would have generated. For example, in a matter where a brokerage firm is alleged to have overcharged its client investors for stock transactions over a period of years, investors may claim that the amount of the overcharge in each transaction would have been reinvested into the same stocks and generated an additional return by the valuation date. In other situations, it is natural to see the past flow as a loss that would have to be financed (e.g., from a bank), and it is necessary to consider what rate of return would have been required in order to finance that loss. For example, in a patent infringement case where the patentee's profits were adversely impacted by the infringing company's conduct, the patentee may claim that its lost profits are economically equivalent to an unsecured loan (albeit an involuntary one) to the infringer, and therefore the infringer's borrowing rate should be used to capitalize past lost profits.

The appropriate capitalization rate to value past flows of economic income or loss is not always a matter of expert analysis. Sometimes, it is set by statute or legal precedent (a topic covered in the section titled "Prejudgment Interest" below), and all an expert needs to do, once the direct economic impact of the alleged harmful act in past periods has been estimated, is to capitalize those estimates through the valuation date.

Disagreements about what valuation date to use

An important decision in many analyses of economic damages is the valuation date, that is, the date as of which the value of the direct economic impact of the alleged harmful act over multiple periods will be determined. Sometimes, the appropriate valuation date is set by law. For example, in a breach of contract matter, the proper valuation date is typically the date of the alleged breach. However, when the valuation date is not set by law, expert disputes may arise over the valuation date, given its potential impact on damages.

There are at least two reasons why the valuation date used in a damages analysis can have a material impact on damages. First, the appropriate rates for discounting and capitalization purposes change significantly over time. For example, all else being equal, the appropriate discount rate to value a business's expected future profits will be higher in periods when future inflation is expected to be high. Thus, if there are significant changes in inflation expectations, the damages valuation date can have a significant impact on the damages estimate. Second, the experts' assessment of the direct economic impact of the alleged harmful act across multiple periods depends on the parties' knowledge as of the valuation date, and that knowledge may change over time as the parties'

economic circumstances evolve. For example, in a business valuation case, the value of the business at issue may be very different at a date when a company has no direct competitors versus a date—potentially just a few months later—after competitor firms have entered the market, significantly eroding the profitability of the business at issue.

Other Issues Arising in General in Damages Measurement

Data Validity

Validation of any dataset is critical. The expert needs to have a firm basis for relying on the chosen data. Opposing parties frequently try to impeach damages estimates by challenging the reliability of the data or an expert's validation of the data. This section discusses some key issues that experts consider when evaluating data validity.

Criteria for Determining the Validity of Data

The validity of data is ultimately a matter of judgment. Experts often need to use data that are not mathematically precise because the only relevant available data may be known to contain some errors. Experts generally have an obligation to use data that are as accurate as possible, meaning that the expert has used every practical means to eliminate erroneous information. Experts should also perform cross-checks with other data, to the extent possible, to demonstrate completeness and reliability. Where data are inherently inaccurate because of random influences, validity requires absence of bias or adjustment for bias. Validation of data turns in part on commonsense indicators of accuracy and bias. The following is a list, in rough order of presumptive validity, of data sources often used in damages measurement:

1. Official government publications and databases, such as from the Census Bureau, the Bureau of Labor Statistics, and the Bureau of Economic Analysis
2. Filings with the Securities and Exchange Commission, including a company's audited financial statements
3. A company's accounting and sales records maintained in the normal course of business
4. A company's operating reports prepared for management in the normal course of business
5. Reports issued by securities analysts covering a company

6. A survey designed by the damages expert with assistance from survey professionals, conforming to established standards of survey design and execution
7. A marketing-research study conforming to established standards for these studies
8. Industry reports and other materials prepared by unaffiliated organizations and consultants
9. Newspaper articles
10. A company's study of damages from the harmful event, prepared in the normal course of business
11. A company's study of damages, prepared for litigation

Other factors can alter this presumptive order of validity. When audited financial statements are alleged to be fraudulent, they (or some portions therein) may lose their presumption of validity. The most fully researched articles in the best newspapers have a higher presumption of validity. Some private industry reports are highly reliable. Rules of evidence may also affect when and how these various data sources can be used in expert reports and at trial. However, when an expert's preferred analysis would rely on data from an opposing party and those data are unavailable, then courts may make allowances for this lack of availability if the expert has made every effort to demonstrate that the data used in the expert's actual analysis are reasonable.

Quantitative Methods for Validation

One important aspect of validation is to verify that the data are complete. If a separate summary document is available that shows the number of records in the database together with summary statistics such as total amounts paid, completeness is easy to establish. Other methods for establishing completeness include examining serial numbers for records and finding other sources of information about transactions that should be in the data and verifying the presence of all or a sample of them.

Another validation method is to examine the accuracy of a sample of observations in a dataset by comparing the sampled observations to corresponding source documents to ensure the values match. Choosing which specific observations to sample can be done in alternative ways. For example, if the dataset consists of purchasing records, then the expert may examine all of the records for a sample of customers, or the expert may examine all of the records for a sample of transactions of certain types.

Another important aspect of validation is to test the internal integrity of the data. For example, sales proceeds should reflect underlying quantities sold and unit prices. The expert can compare sales proceeds to the product of quantities

and prices. Similarly, an expert may analyze the distribution of values for a particular data field to identify observations with values outside the expected range. For example, a negative value in a field representing a unit price would not make logical sense.

Handling of Missing Data

In dealing with missing data, it is critical to ascertain why the data are missing and to attempt to isolate the extent of the missing data. If only a small fraction of data is missing and the pattern appears to be random, then potentially the issue of the missing data can be disregarded and inferences can be drawn using only the available data.

However, missing data are seldom random. For example, suppose that only 1% of transactions are missing, but the transactions that are missing are large ones accounting for a third of all volume. Validation by summarizing across different characteristics will usually identify missing data that are not random. Such errors might occur if all the missing transactions were submitted in a different format that the program for reading the data does not handle. For example, a manufacturer might record sales to Walmart separately from all other customers.

Identifying and adding the missing data to the database is the best correction for missing data, but this is often not possible. In this case, the expert needs to address the problem in another way. The simplest method is to "gross up" damages to reflect the missing data. For example, if 10% of the transactions are randomly missing, then the expert may correct for the missing transactions by dividing calculated damages by 0.9. This method implicitly assumes that the percentage of missing transactions is known and that the missing and nonmissing transactions reflect the same damages on average.

A related approach is to rely on partial, detailed data to measure damages as a fraction of another variable, such as sales. With this approach, other reliable and complete company records such as audited financials may be used to identify the company's total revenues, and damages are then calculated as the fraction of total sales as calculated above. This approach implicitly assumes that the relationship between the missing variable (i.e., damages) and the related, observable variable (e.g., sales) is stable and that most of the variation in outcomes for the missing variable can be explained by the variation in outcomes for the observable variable. The expert may choose to patch together incomplete data from one source and infer complete, reliable data using other approaches. Such approaches can include using a survey of customers or workers to measure damages per dollar of sales or per dollar of earnings and then applying those ratios to reliable data on total sales or total earnings.

Prejudgment Interest

The law may specify how to calculate interest for losses prior to a verdict on liability, generally termed prejudgment interest.[34] The law may exclude prejudgment interest, specify prejudgment interest to be a statutory rate, or exclude compounding.[35] Table 1 illustrates these alternatives. With simple uncompounded interest, losses from five years before trial earn five times the specified annual interest rate, so compensation for a $100 loss from five years ago is $135 at a 7% annual interest rate. With compound interest, the plaintiff earns interest on past interest. Compensation at 7% interest compounded annually is about $140 for a loss of $100 five years before trial. The difference between simple and compound interest becomes much larger if the time from loss to trial is greater or if the interest rate is higher. Because interest is often reinvested to earn further interest, economic analysis generally supports the use of compound interest.

Table 1. Calculations of Prejudgment Interest (in Dollars)

Years of Before Trial	Loss Without Interest	Loss with Compound Interest at 7%	Loss with Simple Uncompounded Interest at 7%
10	$100	$197	$170
9	$100	$184	$163
8	$100	$172	$156
7	$100	$161	$149
6	$100	$150	$142
5	$100	$140	$135
4	$100	$131	$128
3	$100	$123	$121
2	$100	$114	$114
1	$100	$107	$107
0	$100	$100	$100
Total	$1,100	$1,578	$1,485

34. *See generally* Michael S. Knoll, *A Primer on Prejudgment Interest*, 75 Tex. L. Rev. 293 (1996) (discussing prejudgment interest extensively). *See, e.g.*, Ford v. Rigidply Rafters, Inc., 984 F. Supp. 386, 391–92 (D. Md. 1997) (specifying a method of calculating prejudgment interest in an employment discrimination case to ensure plaintiff is fairly compensated rather than given a windfall); Arcon/Pac. Ltd. v. Coit, No. C-81-4264-VRW, 1997 WL 578673, at *1–*2 (N.D. Cal. Sep. 8, 1997) (reviewing supplemental interest calculations and applying California state law to determine the appropriate amount of prejudgment interest to be awarded); Prestige Cas. Co. v. Michigan Mut. Ins. Co., 969 F. Supp. 1029, 1032–34 (E.D. Mich. 1997) (analyzing Michigan state law to determine the appropriate prejudgment interest award).

35. When compounding interest, it is important to use a rate that reflects the periodicity at which interest is compounded (e.g., one should use an annual rate if interest is compounded annually).

Where the law does not prescribe the form of interest for past losses, experts will normally apply a reasonable interest rate to bring those losses forward. The parties may disagree on whether the interest rate should be measured before or after tax. The before-tax interest rate is the normally quoted rate. To calculate the corresponding after-tax rate, one subtracts the amount of income tax the recipient would have to pay on the interest. Thus, the after-tax rate depends on the tax situation of the plaintiff. As an example, if the before-tax interest rate is 9% and the tax rate is 30%, then one would subtract 2.7% (9% times 30%) from the before-tax interest rate of 9% to arrive at an after-tax interest rate of 6.3%.

Even where damages are calculated on a pretax basis, economic considerations suggest that the prejudgment interest rate should be on an after-tax basis: Had a taxpaying plaintiff actually received the lost earnings in the past and invested the earnings at the assumed rate, income tax would have been due on the interest. The plaintiff's accumulated value would be the amount calculated by compounding past losses at the after-tax interest rate.

Where there is economic disparity between the parties, there may be a disagreement about whose interest rate should be used—the borrowing rate of the defendant or the lending rate of the plaintiff or some other rate. There may also be disagreements about adjustment for risk.[36]

Accounting for Income Taxes

A damages award compensates the plaintiff for lost economic value. In principle, the calculation of compensation should measure the plaintiff's loss after taxes and then calculate the magnitude of the pretax award needed to compensate the plaintiff fully, once taxation of the award is considered. In practice, the tax rates applied to the original loss and to the compensation are frequently the same. When the rates are the same, the two tax adjustments are a wash. In that case, the appropriate pretax compensation is simply the pretax loss, and the damages calculation may be simplified by the omission of tax considerations.

In some damages analyses, explicit consideration of taxes is essential, and disagreements between the parties may arise about these tax issues. If the plaintiff's lost income would have been taxed as a capital gain (at a preferential rate), but the damages award will be taxed as ordinary income, the plaintiff can be expected to include an explicit calculation of the extra compensation needed to

36. *See generally* James M. Patell, Roman L. Weil & Mark A. Wolfson, *Accumulating Damages in Litigation: The Roles of Uncertainty and Interest Rates*, 11 J. Legal Stud. 341 (1982), https://doi.org/10.1086/467705 (extensive discussion of interest rates in damages calculations).

make up for the loss of the tax advantage. Sometimes tax considerations are paramount in damages calculations.[37]

Example: Trustee wrongfully sells Beneficiary's property at full market value. Beneficiary would have owned the property until death and deferred the capital gains tax.

Comment: Damages are the difference between the actual capital gains tax and the present value of the future capital gains tax that would have been paid but for the wrongful sale.

In some cases, the law requires different tax treatment of the claimed loss and compensatory awards. Again, the tax adjustments do not offset each other, and consideration of taxes may be a source of dispute.

Example: Driver injures Victim in a truck accident. A state law provides that awards for personal injury are not taxable, even though the income lost as a result of the injury would have been taxable. Victim calculates damages as lost pretax earnings, but Driver calculates damages as lost earnings after tax.[38] Driver argues that the nontaxable award would exceed actual economic loss if it were not adjusted for the taxation of the lost income.

Comment: Under the principle that damages are to restore the plaintiff to the economic equivalent of the plaintiff's position absent the harmful act, it may be recognized that the income to be replaced by the award would have been taxed.[39] However, the law in a particular jurisdiction may not allow a jury instruction on the taxability of an award.[40]

37. *See generally* John H. Derrick, Annotation, *Damages for Breach of Contract as Affected by Income Tax Considerations*, 50 A.L.R. 4th 452 (1986) (discussing a variety of state and federal cases in which courts ruled on the propriety of tax considerations in damage calculations; courts have often been reluctant to award difference in taxes as damages because it is calling for too much speculation).

38. *See generally* Brian C. Brush & Charles H. Breedon, *A Taxonomy for the Treatment of Taxes in Cases Involving Lost Earnings*, 6 J. Legal Econ. 1 (1996) (discussing four general approaches for treating tax consequences in cases involving lost future earnings or earning capacity based on the economic objective and the tax treatment of the lump sum award).

39. *See, e.g.*, Myers v. Griffin-Alexander Drilling Co., 910 F.2d 1252 (5th Cir. 1990) (holding loss of past earnings between the time of the accident and the trial could not be based on pretax earnings).

40. *See generally* John E. Theuman, Annotation, *Propriety of Taking Income Tax into Consideration in Fixing Damages in Personal Injury or Death Action*, 16 A.L.R. 4th 589 (1982) (discussing a variety of state and federal cases in which the propriety of jury instructions regarding tax consequences is at issue). *See, e.g.*, Bussell v. DeWalt Prods. Corp., 519 A.2d 1379 (N.J. 1987) (holding that trial court hearing a personal injury case must instruct jury, upon request, that personal injury damages are not subject to state and federal income taxes); Gorham v. Farmington Motor Inn, Inc., 271 A.2d 94 (Conn. 1970) (holding court did not err in refusing to instruct jury that personal injury damages were tax-free).

Example: Worker is wrongfully deprived of tax-free fringe benefits by Employer. Under applicable law, the award is taxable. Worker's damages estimate is grossed up by a factor so that the amount of the award, after tax, is sufficient to replace the lost tax-free value.

Comment: Again, to achieve the goal of restoring the plaintiff to a position economically equivalent absent the harmful act, an adjustment of this type is appropriate. The adjustment is often called grossing up damages.[41] To accomplish grossing up, divide the lost tax-free value by one minus the tax rate. For example, if the loss is $100,000 of tax-free income, and the income tax rate is 25%, the award should be $100,000 divided by 0.75, or $133,333.

Damages with Multiple Challenged Acts: Disaggregation

Plaintiffs sometimes challenge a number of defendants' acts and offer an estimate of the combined effect of those acts. If the court determines that only some of the challenged acts are illegal, the damages analysis needs to be adjusted to consider only those acts. Ideally, the damages testimony would equip the factfinder to determine damages for any combination of the challenged acts, but that may be tedious. If there are, say, ten challenged acts, it would take more than 1,000 separate studies to determine damages for every possible combination of findings about the unlawfulness of the acts.

There have been several cases where the jury has found partially for the plaintiff, but the jury lacked assistance from the damages experts on how the damages should be calculated for the combination of acts the jury found to be unlawful. Although some of these juries have attempted to resolve these issues, appeals courts have sometimes rejected damages found by juries without supporting expert testimony.[42]

One solution to this problem is to make the determination of the illegal acts before damages testimony is heard, termed bifurcation of liability and damages.

41. *See* Cecil D. Quillen, Jr., *Income, Cash, and Lost Profits Damages Awards in Patent Infringement Cases*, 2 Fed. Cir. Bar J. 201, 207 (1992) (discussing the importance of taking tax consequences and cash flows into account when estimating damages).

42. *See, e.g.*, Litton Sys., Inc. v. Honeywell Inc., Case No. CV 90–4823 MRP (Ex), 1996 U.S. Dist. LEXIS 14662, at *13 (C.D. Cal. July 24, 1996) (granting new trial on damages only "[b]ecause there is no rational basis on which the jury could have reduced Litton's 'lump sum' damage estimate to account for Litton's losses attributable to conduct excluded from the jury's consideration"); Image Tech. Servs., Inc. v. Eastman Kodak Co., 125 F.3d 1195, 1224 (9th Cir. 1997), *cert. denied*, 523 U.S. 1094 (1998) (plaintiffs "must segregate damages attributable to lawful competition from damages attributable to Kodak's monopolizing conduct").

The damages experts can adjust their testimony to consider only the acts found to be illegal.

In some situations, damages are the sum of separate damages for the various illegal acts. For example, there may be one injury in New York and another in Oregon. Then the damages testimony may consider the acts separately, and disaggregation is not challenging.

When the challenged acts have effects that interact, it is not possible to consider damages separately and add up damages for each individual act. This is an area of great confusion. When the harmful acts substitute for each other, the sum of damages attributable to each separately is *less* than their combined effect. As an example, suppose that the defendant has used exclusionary contracts and anticompetitive acquisitions to ruin the plaintiff's business. However, the plaintiff's business could not survive if either the contracts or the acquisitions were found to be legal. Damages for the combination of acts are the value of the business, which would have thrived absent both the contracts and the acquisitions. Now consider damages if only the contracts but not the acquisitions are illegal. In the but-for analysis, the acquisitions are hypothesized to occur because they are not illegal, but not the contracts. But the plaintiff's business cannot function in that but-for situation because the acquisitions alone were sufficient to ruin the business. Hence, damages—the difference in value of the plaintiff's business in the but-for and actual situations—are zero. The same would be true for a separate damages measurement for the acquisitions, with the contracts taken to be legal but not the acquisitions. Thus, the sum of damages for the individual acts is zero, but the damages if both acts are illegal are the value of the business.

When the effects of the challenged conduct are complementary, the sum of damages for each type of conduct by itself will be *more* than damages for all types of conduct together. For example, suppose a party claims that a contract is exclusionary based on the combined effect of the contract's duration and its liquidated damages clause that includes an improper penalty provision. The actual amount of the penalty would cause little exclusion if the duration were brief, but substantial exclusion if the duration were long. Similarly, the actual duration of the contract would cause little exclusion if the penalty were small but substantial exclusion if the penalty were large. A damages analysis for the penalty provision in isolation compares but-for—without the penalty provision but with long duration—to actual, where both the penalty provision and long duration are in effect, and finds that damages are large. Similarly, a damages estimate for the duration in isolation gives large damages. The sum of the two estimates is nearly double the damages from the combined use of both provisions.

A request that the damages expert disaggregate damages among each of the individual challenged acts is therefore complex and will not necessarily lead to individual estimates that will add up to the estimate based on all of the challenged acts. In principle, a separate damages analysis—with its own carefully

specified but-for scenario and analysis—needs to be done for every possible combination of illegal acts.[43]

Example: Hospital challenges Glove Maker for illegally obtaining market power through the use of long-term contracts and the use of a discount program that gives discounts to consortiums of hospitals if they purchase exclusively from Glove Maker. The jury finds that Glove Maker has attempted to monopolize the market with its discount programs but that the long-term contracts were legal because of efficiencies. Hospital states that its damages are the same as in the case in which both acts were unlawful because either act was sufficient to achieve the observed level of market power. Glove Maker argues that damages are zero because the lawful long-term contracts would have been enough to allow it to dominate the market.

Comment: The appropriate damages analysis is based on a careful new comparison of the market with and without the discount program. The but-for analysis should include the presence of the long-term contracts because they were found to be legal.

Disputes About Whether the Plaintiff Is Entitled to All the Damages

When the plaintiff is in some sense a conduit to other parties, the defendant may argue that the plaintiff is entitled to only those damages that it would have retained in the but-for scenario. In the following example, a regulated utility is arguably a conduit to the ratepayers:

Example: Generator Maker overcharges Utility. Generator Maker argues that the overcharge would have been part of Utility's rate base, and so Utility's regulator had set higher prices because of the overcharge. Utility, therefore, did not lose anything from the overcharge. Instead, the ratepayers paid the overcharge. Utility argues that it stands in for all the ratepayers and that the damages

43. Order Denying Motion for Class Certification, Reitman v. Champion Petfoods United States, Inc., No. 18-cv-1736-DOC-JPRx, 2019 U.S. Dist. LEXIS 221941 (C.D. Cal. Oct. 30, 2019). In denying class certification the court found that "the [Plaintiffs' damages] model does not test whether the corrective statements have a different impact on consumers in particular combinations. Therefore, if any theory of liability is disposed of, the survey does not provide a reliable way to determine the difference in value based on the remaining theories of liability because the expert did not test the impact of particular combinations of statements. For these reasons, Plaintiffs have not 'show[n] that such damages can be determined without excessive difficulty and attributed to their theory of liability.'" *Id.* at *32 (citing JustFilm, Inc. v. Buono, 847 F.3d 1108, 1121 (9th Cir. 2017)).

award will accrue to the ratepayers by the same principle—the regulator will set lower rates because the award will count as revenue for rate-making purposes.

Comment: In addition to the legal issue of whether Utility does stand in for ratepayers, there are two factual issues: Was the overcharge actually passed on to ratepayers? Will the award be passed on to ratepayers?

Similar issues can arise in the context of employment law.

Example: Plaintiff Sales Representative sues for wrongful denial of a commission. Sales Representative has subcontracted with another individual to do the actual selling and pays a portion of any commission to that individual as compensation. The subcontractor is not a party to the suit. Defendant Manufacturer argues that damages should be Sales Representative's lost profit measured as the commission less costs, including the payout to the subcontractor. Sales Representative argues that they are entitled to the entire commission.

Comment: Given that the subcontractor is not a plaintiff, and Sales Representative avoided the subcontractor's commission, the literal application of standard principles for damages measurement would appear to call for the lost-profit measure. The subcontractor, however, may be able to claim their share of the damages award. In that case, damages would equal the entire lost commission, so that, after paying off the subcontractor, Sales Representative receives exactly what they would have received absent the breach.

Limitations on Damages

Uncertainty and Speculation in Damages Estimates

Generally, a plaintiff may not recover damages beyond an amount proven with reasonable certainty.[44] This rule permits damages estimates that are not mathematically certain but excludes those that are speculative.[45] Failure to prove damages to a reasonable certainty is a common defense.[46]

44. *See, e.g.*, Restatement (Second) of Contracts § 352 (1981) ("Damages are not recoverable for loss beyond an amount that the evidence permits to be established with reasonable certainty.").

45. Restatement (Second) of Contracts § 352 cmt. a states, in pertinent part: "Damages need not be calculable with mathematical accuracy and are often at best approximate."

46. *See, e.g.*, Cole v. Homier Distrib. Co., 599 F.3d 856, 866 (8th Cir. 2010) (expert testimony on lost profits excluded under *Daubert* standard because it "failed to rise above the level of speculation."). *See also* Webb v. Braswell, 930 So. 2d 387 (Miss. 2006). In *Webb*, the plaintiff's expert

However, defining what constitutes reasonable certainty or speculation in a damages analysis may be difficult. The exclusion of damages on grounds of excessive uncertainty regarding the amount of damages may result in an award of zero damages—even when it is likely that the plaintiff suffered significant damages, though the actual amount is quite uncertain.

Traditionally, damages are calculated without reference to uncertainty about outcomes in the but-for scenario. Outcomes are taken as actually occurring if they are the expected outcome and as not occurring if they are not expected to occur. This approach may overcompensate some plaintiffs and undercompensate others. For example, suppose that a drug company was deprived of the opportunity to bring to market a drug that had a 90% chance of receiving Food and Drug Administration (FDA) approval, at a profit of $2 billion, and a 10% chance of not receiving FDA approval, with losses of $1 billion. The court may treat 90% as near enough to certainty and ignore the 10% risk of failure and award damages of $2 billion.

By contrast, economists quantify losses from uncertain outcomes in terms of expected values, where the value in each outcome is weighted by its probability. Under that approach, economic losses in the above example would be calculated as the foregone $2 billion potential profit assuming FDA approval times 90% plus the $1 billion potential loss times 10%, or $(0.9) \times (\$2 \text{ billion}) + (0.1) \times (-\$1 \text{ billion}) = \$1.7$ billion. Thus, the plaintiff would be overcompensated by $300 million under the traditional approach that ignored the small probability of failure.

Now suppose the drug has only a 40% chance of FDA approval with the same economic payoffs. Although the court may decide to award the plaintiff no damages on grounds of uncertainty and speculation, the economic loss would be calculated as $(0.4) \times (\$2 \text{ billion}) + (0.6) \times (-\$1 \text{ billion}) = \$200$ million. This issue can arise in any situation involving businesses or products whose future profitability is highly uncertain due to an unproven track record of profitability.

Damages That Are Too Remote

A second limitation on damages is that a plaintiff may not recover damages that are too remote. In tort cases, this restriction is expressed in terms of proximate cause,[47] which often is equivalent to reasonable foreseeability. In contract cases,

sought to testify as to future damages resulting from unplanted crops, without establishing that the crops would have been profitable. The court excluded the testimony based on Mississippi's adoption of the *Daubert* standard, stating that "damages for breach of contract must be proven with reasonable certainty and not based merely on speculation and conjecture." *Id.* at 397.

47. *See* William L. Prosser, Palsgraf *Revisited*, 52 Mich. L. Rev. 1 (1953); Osborne M. Reynolds, Jr., *Limits on Negligence Liability:* Palsgraf *at 50*, 32 Okla. L. Rev. 63 (1979).

the limitation is similarly embodied in the idea of reasonable foreseeability—a party may not recover damages that were not reasonably foreseeable by the parties at the time of the agreement.[48] Generally, a party is liable for what are known as direct or general damages—those damages that arise naturally from the breach itself. A defendant may also be liable for consequential or special damages—damages apart from those arising naturally from the breach—if such damages were reasonably foreseeable at the time of the agreement.[49]

Mitigation of Losses

A third limitation on damages is that a party may not recover for losses it could have avoided; this limitation is often expressed by stating that the injured party has a duty to mitigate, or lessen, its damages. The economic justification for the mitigation rule is that the injured party should not cause economic waste by needlessly increasing its losses.[50]

For example, in a dispute about mitigation, a defendant may propose that any damages estimate should include an offset for earnings the plaintiff should have achieved, under proper mitigation, rather than actual earnings. As another example, the defendant might presume that the plaintiff could mitigate by locating another source of supply in the event of a breach of a supply agreement. Damages are limited to the difference between the contract price and the current market price in that situation.

For personal injuries, the issue of mitigation often arises because the defendant believes that the plaintiff's failure to work after the injury is a withdrawal from the labor force or retirement rather than the result of the injury.[51] For commercial torts, mitigation issues can be more subtle. Where the plaintiff believes that the harmful act destroyed a company, the defendant may argue that the company could have been put back together and earned profit, possibly in a different line of business.[52] The defendant will then treat the hypothetical profits as an offset to damages.[53]

Alternatively, where the plaintiff continues to operate the business after the harmful act and includes subsequent losses in damages, the defendant may argue that the proper mitigation was to shut down after the harmful act.[54]

48. *See* E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1199–1210.

49. *Id.*

50. *See id.* at 1183–84.

51. *See* William T. Paulk, *Mitigation Through Employment in Personal Injury Cases: The Application of the "Reasonable" Standard and the Wealth Effects of Remedies,* 58 Ala. L. Rev. 647–64 (2007).

52. *See* Seahorse Marine Supplies Inc. v. Puerto Rico Sun Oil Co., 295 F.3d 68, 84–85 (1st Cir. 2002).

53. *Id.* at 84.

54. *Id.* at 85. *See also In re* First New England Dental Ctrs., Inc., 291 B.R. 229, 240 (D. Mass. 2003).

Example: Franchisee Soil Tester starts up a business based on Franchiser's proprietary technology, which Franchiser represents as meeting government standards. During the startup phase, Franchiser notifies Soil Tester that the technology has failed. Soil Tester continues to develop the business but sues Franchiser for profits it would have made from successful technology. Franchiser calculates much lower damages on the theory that Soil Tester should have mitigated by terminating the startup.

Disagreements about mitigation may be hidden within the frameworks of the plaintiff's and the defendant's damages studies.

Example: Defendant Board Maker has breached an agreement to supply circuit boards. Plaintiff Computer Maker's damages study is based on the loss of profits on the computers to be made from the circuit boards. Board Maker's damages study is based on the difference between the contract price for the boards and the market price at the time of the breach.

Comment: There is an implicit disagreement about Computer Maker's duty to mitigate by locating alternative sources for the boards not supplied by the defendant. The Uniform Commercial Code spells out the principles for resolving these legal issues under the contracts it governs.[55]

Damages That May Exceed the Cost of Avoidance

An important consideration in capping damages may be the costs of steps that the plaintiff could have taken that would have eliminated damages. This argument is closely related to mitigation but has an important difference: The defendant may argue that the plaintiff's failure to undertake a costly step that would have avoided losses was reasonable, but that the failure to take that step shows that the plaintiff knew that damages were much smaller than its later damages claim.

55. *See, e.g.*, Aircraft Guar. Corp. v. Strato-Lift, Inc., 991 F. Supp. 735, 738–39 (E.D. Pa. 1998) (Mem.) (finding that according to the Uniform Commercial Code, plaintiff-buyer had a duty to mitigate if the duty was reasonable in light of all the facts and circumstances, but that failure to mitigate does not preclude recovery); S.J. Groves & Sons Co. v. Warner Co., 576 F.2d 524 (3d Cir. 1978) (holding that the duty to mitigate is a tool to lessen plaintiff's recovery and is a question of fact); Thomas Creek Lumber & Log Co. v. United States, 36 Fed. Cl. 220 (1996) (finding that under federal common law, the U.S. government had a duty to mitigate in breach-of-contract cases).

Example: Insurance Company suffered a business interruption because a fire made its offices unusable for a period of time. Insurance Company's damages claim for $10 million includes not only the lost business until the offices were usable but also damages for permanent loss of business from customers who found other sources during the period the offices were unusable. Defendant argues that the plaintiff's failure to relocate to temporary quarters shows that their losses were less than the $350,000 cost of that relocation.

Comment: Defendant's argument has the unstated premise that Insurance Company could have carried on its business and avoided any of its later losses by relocating. Insurance Company will likely argue that a decision not to relocate was commercially appropriate because relocation would not have avoided much of the lost business.

The Effect of a Liquidated Damages Clause

In addition to legally imposed limitations on damages, the parties themselves may have agreed to impose limits on damages, should a dispute arise. Such clauses are common in many types of agreements. Once litigation has begun, the parties may dispute whether these provisions are legally enforceable. The law may limit enforcement of liquidated damages provisions to those that bear a reasonable relation to the actual damages. In particular, the defendant may attack the amount of liquidated damages as an unenforceable penalty. The parties may disagree on whether the harmful event falls within the class intended by the contract provision.

Changes in economic conditions may be an important source of disagreement about the reasonableness of a liquidated damages provision. One party may seek to overturn a liquidated damages provision on the grounds that new conditions make it unreasonable.

Example: Scrap Iron Supplier breaches supply agreement and pays only the specified liquidated damages. Buyer seeks to set aside the liquidated damages provision because the price of scrap iron has risen, and the liquidated damages are a small fraction of actual damages under the expectation principle.

Comment: There may be conflict between the date for judging the reasonableness of a liquidated damages provision and the date for measuring expectation damages, as in this example. Generally, the date for evaluating the reasonableness of liquidated damages is the date the contract is made. In contrast, the date for measuring

expectation damages is the date of the breach. The conflict may be resolved by the substantive law of the jurisdiction.

Damages in Class Actions

Class Certification

The U.S. Supreme Court's decision in *Comcast Corp. v. Behrend* states that "rigorous analysis" is required to establish if the model proposed by plaintiffs in a class action is capable of measuring damages on a class-wide basis.[56] Consistent with this requirement, courts commonly analyze damages models as part of their decision whether to certify a class.[57]

A key question often analyzed by courts is whether the proposed method of measuring class-wide damages isolates the impact of the theory of harm alleged. In *Comcast*, the Supreme Court found that the lower court erred in certifying a class of cable television subscribers because the plaintiffs had failed to show that the proposed damages model adequately translated the legal theory of the harmful event into an analysis of the economic impact of that event. Specifically, the plaintiffs proposed a damages model that estimated the combined impact of four proposed theories of antitrust impact but failed to estimate damages specific to each theory. Because the lower court eventually dismissed all but one of the plaintiffs' theories of harm and the plaintiffs' expert acknowledged that his model did not isolate damages resulting from any one of the plaintiff's theories of harm, the Supreme Court found that the proposed damages model was inadequate in that it was unable to isolate the harm resulting from the single surviving theory.[58]

Since *Comcast*, many courts have emphasized the importance of having a damages model that ties directly to the theory of harm. For example, in *Reitman v. Champion*, a matter involving alleged misrepresentations on the packaging of dog food, the court denied class certification because the plaintiffs' damages

56. 569 U.S. 27, 35 (2013).

57. *See, e.g.*, Memorandum of Law & Order, Johnson v. Evangelical Lutheran Church in Am., No. 11–00023 (MJD/LIB), 2013 U.S. Dist. LEXIS 41944 (D. Minn. Mar. 26, 2013) (court evaluated if Board of Pensions of the Evangelical Lutheran Church in America breached its fiduciary duties by reducing annuity payments and rejected class certification on the basis that the board actions helped rather than harmed the majority of putative class members); Memorandum Opinion and Order, Indiana Pub. Requirement Sys. v. AAC Holdings, Inc., No. 3:19-cv-00407, 2023 U.S. Dist. LEXIS 31022 (M.D. Tenn. Feb. 24, 2023) (The court evaluated whether plaintiff had put forth a reliable method to estimate damages consistent with the alleged misrepresentations. The court rejected plaintiff's damages model with respect to plaintiff's materialization-of-the-risk theory because it did not factor in the risk on which the theory was based. The court accepted plaintiff's damages model otherwise.).

58. *Comcast*, 569 U.S. at 38.

methodology tested the effects of statements created by plaintiffs to correct the alleged misrepresentations, and not the effects of the alleged misrepresentations on consumers.[59] Similarly, in *Freitas v. Cricket Wireless*, a case involving the sale of 4G-capable phones in markets where the defendant did not actually provide 4G coverage, the court denied class certification because plaintiffs simply measured the price difference between various 3G and 4G phones, but failed to account for other differences in the phones, such as the brand and features of the phones.[60]

Another question that courts commonly analyze in connection with the requirements for class certification is whether the damages model proposed by the plaintiffs results in the need for individualized inquiry.[61] Courts have usually accepted damages models that offer a common or formulaic method to estimate damages even if those models result in individualized damages calculations.[62] However, class certification may be rejected in cases where computing individual damages is deemed overly complex or fact-specific.[63]

A third common question is whether the proposed damages methodology is sufficiently developed for the court to assess if the model is capable of measuring class-wide damages. Courts have rejected models that fail to explain how important variables will be accounted for,[64] that do not plausibly explain how the consequences of the allegations will be isolated,[65] or that are "vague, indefinite, and unspecific."[66] At the same time, some courts have accepted damages models that have not been implemented but merely proposed at an adequate level of detail.[67]

59. The court also indicated that "if any theory of liability is disposed of, [the plaintiffs' damages model] does not provide a reliable way to determine the difference in value based on the remaining theories of liability because the expert did not test the impact of particular combinations of statements." *See* Order Denying Motion for Class Certification, Reitman v. Champion Petfoods USA, Inc., No. 18-cv-1736-DOC-JPRx, 2019 U.S. Dist. LEXIS 221941 (C.D. Cal. Oct. 30, 2019).

60. Freitas v. Cricket Wireless, LLC, No. C 19–07270 WHA, 2022 WL 3018061, at *6 (N.D. Cal. July 29, 2022).

61. *See, e.g.*, Ludlow v. BP, P.L.C., 800 F.3d 674 (5th Cir. 2015).

62. *See, e.g.*, *In re* Whirlpool Corp., 722 F.3d 838 (6th Cir. 2013); Leyva v. Medline Indus., Inc., 716 F.3d 510 (9th Cir. 2013).

63. In *Behrend v. Comcast Corp.*, the court reiterated that class-certification damages should not require "labyrinthine individual calculations." 655 F.3d 182, 206 (3d Cir. 2011). *See also* Brown v. Electrolux Home Prods., Inc., 817 F.3d 1225 (11th Cir. 2016) (stating "individual damages defeat predominance if computing them 'will be so complex, fact-specific, and difficult that the burden on the court system would be simply intolerable.'") (citation omitted).

64. Bruton v. Gerber Prods. Co., No. 12-cv-02412-LHK, 2018 U.S. Dist. LEXIS 30814, at *34 (N.D. Cal. Feb. 13, 2018).

65. Ang v. Bimbo Bakeries USA, Inc., No. 13-cv-01196-HSG, 2018 U.S. Dist. LEXIS 149395, at *2 (N.D. Cal. Aug. 31, 2018).

66. *See also* Ohio Pub. Emps. Ret. Sys. v. Fed. Home Loan Mortg. Corp., No. 4:08CV0160, 2018 WL 3861840 (N.D. Ohio Aug. 14, 2018).

67. *See, e.g.*, *In re* Scotts EZ Seed Litig., 304 F.R.D. 397 (S.D.N.Y. 2015). *See also* Goldemberg v. Johnson & Johnson Consumer Cos., 317 F.R.D. 374 (S.D.N.Y. 2016).

A court operating under the rule that class certification requires a fully developed damages quantification may need to grant discovery prior to class certification to support the class's damages analysis and the defendant's opposition.

Finally, as in other types of litigation, courts have evaluated damages models proposed in class actions with respect to their relevance and reliability. Many courts have rejected damages models proposed in class actions because they fail to meet the standard of relevance and reliability set by *Daubert*.[68]

Class-Wide Damages

In many class actions, the damages expert measures and testifies to class-wide damages using methods discussed elsewhere in this reference guide. At a high level, the damages method offered by the expert may take any of a number of forms. Among other possibilities, it may involve a uniform damages amount per class member, or per instance of harm, multiplied by the total number of class members or instances of harm; it may involve a uniform damages amount per time period (e.g., workweek) multiplied by the number of at-issue time periods; it may involve a uniform damages percentage (e.g., an overpayment percentage) that is applied to total sales; or it may involve the application of a common method (e.g., a formula) to calculate differing damages amounts for different class members based on various inputs.

In class actions involving securities, damages experts are usually required to estimate per-share damages. Class-wide damages are typically calculated in a subsequent claims process where individual shareholders submit claims, along with evidence of their trades in the securities at issue.[69]

Finally, courts do not typically require class-wide damages to be measured with certainty. As the Supreme Court stated, "Calculations need not be exact."[70] While mere speculation or guess is not sufficient, courts have held that it is enough if class-wide damages are demonstrated "as a matter of just and reasonable inference."[71]

68. *See, e.g.*, IBEW Local 90 Pension Fund v. Deutsche Bank AG, No. 11 Civ. 4209 KBF, 2013 WL 5815472 (S.D.N.Y. Oct. 29, 2013); *In re* Blood Reagents Antitrust Litig., 783 F.3d 183 (3d Cir. 2015). Note, however, that there is disagreement across district courts about whether the *Daubert* standard should be applied to evidence presented at the class-certification stage. *See, e.g.*, Cox v. Zurn Pex, Plumbing Prods. Liab. Litig., 644 F.3d 604 (8th Cir. 2011); Sali v. Corona Reg'l Med. Ctr., 909 F.3d 996, 1005–06 (9th Cir. 2018).

69. *See, e.g.*, Order Regarding Plaintiffs' Motions for Prejudgment Interest, Approval of Notice of Verdict and Claims Administration Procedure, and Unsealing Documents (Dkt. Nos. 745, 748, 752), Hsingching Hsu v. Puma Biotechnology, Inc., No. SACV 15–00865 AG (SHKx), 2019 U.S. Dist. LEXIS 154278 (C.D. Cal. Sep. 9, 2019).

70. Comcast Corp. v. Behrend, 569 U.S. 27, 35 (2013).

71. *See, e.g.*, Behrend v. Comcast Corp., 655 F.3d 182, 203–04 (3d Cir. 2011) (quoting Story Parchment Co. v. Paterson Parchment Paper Co., 282 U.S. 555, 563 (1931)). *See also In re* MyFord Touch Consumer Litig., 291 F. Supp. 3d 936 (N.D. Cal. 2018).

Damages of Individual Class Members

Damages experts sometimes have a role in the process of disbursing funds from a verdict or settlement to individual class members. For example, an expert can develop software that measures individual damages based on evidence supplied by individuals through a claims-processing facility. In *Millsap v. McDonnell Douglas*, more than 1,000 class members—persons affected by the defendant's challenged layoffs—completed sworn questionnaires describing their post-layoff experiences.[72] McDonnell Douglas supplied additional information from its employee records. The class's damages expert used standard methods for valuing claims for lost earnings to calculate estimates for each class member. Because the settlement reflected a compromise among the parties on a number of disputes about the law and about the facts underlying the layoffs, the total cash from the settlement was less than the sum of these estimates, and class members received a fraction of the amount indicated by the damages model.

Damages as a Basis to Evaluate the Fairness of a Proposed Settlement

Class-wide damages measures have a key role in resolving class-action cases, because courts refer to them in determining the fairness of proposed settlements. The court's careful review of the benefits proposed for the class is essential because the interests of the class's counsel may not be aligned with those of the class members with respect to settlement.

Example: Lender required excessive escrow deposits for property taxes from a class of mortgage borrowers, although the excess was repaid at the end of each year. Under the terms of the proposed settlement negotiated between Lender's lawyers and those for the class, the excess is to be refunded to the class members immediately, with 30% of that amount paid to class counsel as fees.

Comment: This settlement is unreasonable and leaves the class worse off than they were under the excessive escrows. The loss to the class from placing funds in an escrow is the foregone interest on the amount in the escrow, which would likely be no more than 10% of the excess amount of the escrow. By granting 30% of the refund as fees to class counsel, the class members are at least 20% worse off than they would be if the excess were repaid with a delay.

72. No. 94-cv-633-H(M), 2003 WL 21277124 (N.D. Okla. May 28, 2003).

Damages Issues in Selected Types of Cases

Estimation of Economic Damages in Consumer Product Liability and Misrepresentation Cases

Consider consumer class actions in which a product is alleged to have a particular defect that the plaintiff claims should have been disclosed to consumers, or in which a product is alleged to have been misrepresented to consumers before they purchased it. In these types of cases, a common damages claim is that consumers would not have bought the product, or would have paid less for it, had they known about the alleged defect or had the product not been misrepresented.[73] The following sections discuss issues that often appear in the calculation of damages associated with these claims.

Defining the Plaintiff's Economic Position in the Actual and But-For Worlds

As explained earlier in this reference guide, the role of the damages expert is to translate the plaintiff's theory of harm into an objective model for damages. In the case of overpayment damages in a product liability or misrepresentation case, the plaintiff's actual economic position is usually defined as the price paid for the allegedly defective product at a then-current market price without the benefit of appropriate disclosures from the manufacturer.[74] Plaintiffs commonly offer invoices, sales receipts, or other retail sales data as evidence of purchase and of the prices paid. Disputes may arise if the plaintiff cannot offer direct evidence of the quantity of products purchased or the price paid for them (for example, if the manufacturer has data only on sales to distributors), or if the price paid was different across buyers due to individual discounts or negotiations.

The plaintiff's position in the but-for world is generally more complicated to define. In cases where specific statements made by the product manufacturer allegedly misrepresent the product (for example, by including incorrect information on

73. *See, e.g.*, *In re* MyFord Touch Consumer Litig., No. 13-cv-3072-EMC (N.D. Cal. Oct. 13, 2015); First Amended Class Action Complaint for Damages, Zakaria v. Gerber Prods. Co., No. 2:15-cv-0200-JAK (Ex), 2015 WL 1085581 (C.D. Cal. Feb. 27, 2015). Other common damages claims are that the manufacturer profited unfairly from the sale of the defective product and that the plaintiffs suffered economic losses in their attempt to repair the defect. Issues arising from these claims can be analyzed using the frameworks provided in the preceding sections.

74. *See, e.g.*, Hendricks v. Starkist Co., No. 13-cv-00729-HSG, 2016 U.S. Dist. LEXIS 134872 (N.D. Cal. Sep. 29, 2016); Elkies v. Johnson & Johnson Servs., No. 2:17-cv-07320-GW-JEMx, 2020 U.S. Dist. LEXIS 256341 (C.D. Cal. June 22, 2020).

the product label), a standard approach is to assume that in the but-for world such misrepresentations would have not occurred (for example, assuming that the label would not include the allegedly misleading information).[75] Cases in which the plaintiff alleges that the manufacturer omitted relevant information that would have been material for consumers can be more challenging for the damages expert because the characteristics of the disclosure (e.g., wording, size) in the hypothetical but-for world can have significant implications for how consumers react to it and the market price they pay.[76] Moreover, if the alleged defect only manifests in some products with a given probability,[77] the specific way in which information about that probability is conveyed to consumers can also have implications for how they react to it and the market price they pay.[78]

Determining the But-For Price

In some cases, the market price of the product, adjusted to reflect the effect of the allegations, may be an appropriate estimate of the but-for price. For example, if the manufacturer of the product later added a warning to the product that corrected the alleged omissions in a way that is consistent with plaintiffs' allegations, the market price of the product in the period after the warning was added may be an appropriate measure of the price of the product in the but-for world. In this case, damages experts must account for other factors beyond the addition of the warning that may have affected the price of the product over the same period in order to isolate the effect of the alleged misconduct. For example, changes in supply or demand conditions after a warning label was added to a product may lead to changes in the price of the product that are unrelated to the alleged misconduct.

Damages experts may also rely on data from ostensibly similar products or assets to account for the effect of confounding factors. Empirical methods such as regression analysis can use information from similar products to account for confounding factors such as market-wide trends, geographic variation, and

75. *See, e.g.*, Jones v. ConAgra Foods Inc., No. C 12–01633 CRB, 2014 U.S. Dist. LEXIS 81292 (N.D. Cal. June 13, 2014); Larsen v. Vizio, Inc., No. 8:14-cv-01865-CJC-JCG (C.D. Cal. May 5, 2015).

76. For example, a warning label with negative language about the product could lead to a very different outcome compared to neutral language. *See, e.g.*, Christopher M. Heaps & Tracy B. Henley, *Language Matters: Wording Considerations in Hazard Perception and Warning Comprehension*, 133 J. Psych. 341 (1999), https://doi.org/10.1080/00223989909599747.

77. *See, e.g.*, *In re* Takata Airbag Prod. Liab. Litig., 1:14-cv-24009-FAM (1787) (S.D. Fla. Apr. 24, 2021).

78. Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 Econometrica 263 (1979), https://doi.org/10.2307/1914185.

variation in product features, among others.[79] For these types of analyses, experts need to demonstrate that the benchmarks used are sufficiently similar to the product at issue and that the model adequately captures the most important features of a product.[80] Experts also need to show that the prices of benchmark products are affected by similar factors as the price of the product at issue, other than the alleged misconduct (i.e., the prices have "parallel trends").[81]

Some damages experts in product liability and misrepresentation class actions have not relied on observed market prices but instead have relied on stated preference models, which use surveys to try to assess the impact of the alleged omissions or misrepresentations on consumer demand.[82] Some plaintiff experts have relied directly on the results from stated preference models to estimate per-unit overpayment damages, while other experts have combined estimates of impact on consumer demand with models or assumptions about the supply side of the market to generate estimates of but-for prices.

Conjoint Analysis and Supply-Side Considerations

A stated preference model commonly used by experts for plaintiffs is based on a methodology called conjoint analysis.[83] Conjoint analysis is a survey-based methodology used to analyze consumer preferences for products and product features.[84] Conjoint analysis assumes that the subjective value consumers perceive from a product is the sum of the benefits they derive from individual product features or attributes.[85] In a "choice-based" conjoint analysis (the type most often

79. *See, e.g.*, Mark Campbell et al., A.B.A., *Damages in Consumer Class Actions*, *in* A Practitioner's Guide to Class Actions (Marcy Hogan Greer & Amir Nassihi eds., 3d ed. 2021).

80. *See, e.g.*, Kelly C. Bishop et al., *Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality*, 14 Rev. Env't Econ. & Pol'y 260 (2020), https://doi.org/10.1093/reep/reaa001.

81. Joshua D. Angrist & Jörn-Steffen Pischke, Mostly Harmless Econometrics: An Empiricist's Companion (2008).

82. *See* the section titled "Realistic Modeling of Preferences, Incentives, and Constraints of Buyers and Sellers" above.

83. For applications in the automotive industry, see, e.g., *In re Gen. Motors LLC Ignition Switch Litig.*, No.1:14-MD-02543-JMF (S.D.N.Y.); *In re Chrysler-Dodge-Jeep Ecodiesel Mktg., Sales Pracs. & Prods. Liab. Litig.*, No. 3:17-md-02777- EMC (N.D. Cal.). For applications relating to food products, see, e.g., *Zakaria v. Gerber Prods. Co.*, No. 2:15-cv-0200-JAK, 2015 WL 1085581 (C.D. Cal.); *Colangelo v. Champion Petfoods USA, Inc.*, No. 6:18-cv-01228-LEK-DEP, 2020 WL 777462 (N.D.N.Y). For an application in the consumer electronics industry, see *Davidson v. Apple, Inc.*, No. 5:16-cv-4942-LHK, 2018 WL 10716622 (N.D. Cal.).

84. Vithala R. Rao, Applied Conjoint Analysis (2014).

85. For example, an expert could assume that the subjective value that a consumer obtains from buying a TV is equal to the value they derive from the TV having a 55-inch screen, plus the value of having an LED screen, plus the value of having smart features, minus the value sacrificed by paying the TV's price.

used in litigation), the researcher uses a survey to ask consumers to choose which product they prefer among hypothetical products that vary across product attributes. The researcher then analyzes respondents' choices using statistical methods to generate an estimate of the monetary value that a consumer places on different attributes (also known as willingness to pay).[86] A conjoint analysis used in a product liability case will include one or more attributes that relate to the plaintiffs' allegations. For example, if the plaintiffs claim that the frame-refresh rate of a TV was overstated, the conjoint survey could include some hypothetical product options with the frame-refresh rate advertised to consumers, and other options with the (allegedly lower) correct frame-refresh rate. Based on respondents' answers (and the assumptions underlying the conjoint analysis methodology), an expert may be able to estimate the price that a consumer would be willing to pay for a TV with the advertised frame-refresh rate over a TV with the lower refresh rate.

An expert relying on a conjoint analysis should demonstrate that it provides a sufficiently realistic representation of the preferences of the buyers of the product. This may involve demonstrating that the model includes an appropriate set of attributes that define the demand for the product, that respondents are able to choose not to buy any of the options shown to them (known as the no-purchase option), that the predictions generated by the model are reliable, and that the calculation of respondents' willingness to pay is consistent with the assumptions underlying the conjoint methodology.[87] When implemented correctly and using a representative sample of respondents, an expert could, for example, rely on conjoint analysis to estimate the demand for TVs with and without the misrepresentation at issue.[88]

As explained in the section titled "Issues Arising in Connection with the Use of Models to Simulate the But-For Price" above, because market prices are determined by the interaction of demand and supply, consumer-survey–based methods like conjoint analysis by themselves are insufficient to estimate the market price of the product at issue in the but-for world. Experts using conjoint analysis have relied on at least two different assumptions to attempt to account for supply-side factors. Some experts have assumed that the quantity supplied in the but-for world is the same as the quantity supplied in the actual world.[89] The but-for price obtained with this method will be the price necessary to sell the

86. *See, e.g.*, Moshe Ben-Akiva, Daniel McFadden & Kenneth Train, *Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-Based Conjoint Analysis*, 10 Founds. & Trends Econometrics 1 (2019), http://dx.doi.org/10.1561/0800000036.

87. John R. Hauser & Vithala R. Rao, *Conjoint Analysis, Related Modeling, and Applications*, *in* Marketing Research and Modeling: Progress and Prospects 141–68 (Yoram (Jerry) Wind & Paul E. Green eds., 2004). *See also* Rao, *supra* note 84.

88. *See* Ben-Akiva, *supra* note 86.

89. *See, e.g.*, *In re* Dial Complete Mktg. & Sales Pracs. Litig., 320 F.R.D. 326 (D.N.H. 2017); *In re* NJOY, Inc. Consumer Class Action Litig., No. 14-cv-00428-MMM-JEMx, 2015 WL 12732461 (C.D. Cal. May 27, 2015).

historical quantity of product in the but-for world, in which some consumers would potentially have lower willingness to pay for the but-for product. From an economic perspective, the assumption of a constant or "fixed" quantity supplied results in an estimate of damages based solely on a downward shift in demand that the plaintiffs contend would occur in the but-for world, measured at the quantity sold in the actual world.[90] Other experts have assumed that the quantity sold in the but-for world may be different from the quantity sold in the actual world.[91] For example, some experts have estimated a but-for price based on a market simulation in which supply-side factors include hypothetical sellers that, along with the defendant, maximize their individual profits taking into account their competitors' reactions.[92] The validity of the estimates from this method depends on how accurately the assumptions and characteristics of the market simulation reflect the characteristics of the actual market.

Estimation of Economic Damages in Securities Cases

Consider the typical securities case,[93] with a plaintiff investor (often, on behalf of a class of investors) who claims to have suffered damages from misrepresentations (false or misleading statements or omissions) made by a defendant (or group of defendants) in connection with the purchase of a company's stock.[94] In such cases, the plaintiff claims to have suffered investment losses upon the revelation to the market of the false or misleading nature of the misrepresentations through one or more "corrective" disclosures or events. This section introduces key economic issues and areas of disagreement among experts in the analysis of damages in such cases.[95]

90. For this reason, the but-for price calculated using this assumption may not be a market price that would result from the interaction of supply and demand in the relevant market.

91. *See, e.g.*, Earl v. Boeing Co., 611 F. Supp. 3d 345 (E.D. Tex. 2020).

92. *See, e.g.*, Greg M. Allenby et al., *Valuation of Patented Product Features*, 57 J.L. & Econ. 629, 630 (2014), https://doi.org/10.1086/677071.

93. Statutes under which plaintiffs in securities cases may seek to recover damages include, but are not limited to, the Securities Exchange Act of 1934 § 10(b) (and Rule 10b-5 promulgated thereunder) ("Rule 10b-5" cases) and the Securities Act of 1933 § 11 or § 12(a)(2) ("Section 11" or "Section 12" cases).

94. The concepts discussed herein also apply in securities cases where securities other than common stock (e.g., options and bonds) are at issue.

95. As a threshold matter, given the nature of the plaintiff's claims, analysis of damages in securities cases requires that experts have extensive training in financial economics, the discipline within economics that studies security prices and their determinants. Such training is often obtained through doctoral degrees in economics, finance, or accounting, followed by additional experience (in an academic or professional setting) in the analysis of the determinants of security prices and values. Depending on the circumstances, additional training may further bolster the

Inflation and Losses Caused

The economic issues that arise in the analysis of damages in securities cases relate to the concepts of inflation and losses caused. This section introduces these two concepts and how they relate to the measurement of damages in securities cases.

The concept of stock price inflation arises, for example, when analyzing damages in Rule 10b-5 cases, where the "out-of-pocket" measure of damages (calculated as the difference between inflation at the time of purchase and inflation at the time of sale)[96] limits the amount of per-share damages that plaintiffs can recover.[97] Inflation at a given point in time is the difference between (1) the actual price at which the stock then traded and (2) the but-for price that would have existed in the absence of the misrepresentations (also known as the "true" value).[98]

The actual stock price is easily ascertained based on the plaintiff's trading records or market data. However, estimation of the but-for price—and thus, inflation—requires two analytical steps. The first step is a careful delineation of the information that the plaintiff claims the defendants could and should have revealed instead of or in addition to the misrepresentations prior to the plaintiff's purchase (the "but-for disclosure" or "truthful substitute"). This delineation is based on the plaintiff's allegations and typically not on economic analysis. However, this delineation is critical because the second step, which is based on economic analysis, is the application of tools and techniques from financial economics to estimate the stock's but-for price given the but-for disclosure.

The concept of losses caused arises, for example, when analyzing damages in Rule 10b-5, Section 11, and Section 12(a)(2) cases, where the plaintiff's recoverable

expert's qualification to analyze issues at hand (e.g., accounting training may be helpful in securities cases where the misrepresentations relate to a company's accounting disclosures and practices).

This section focuses on the analysis of damages on a per-share basis. Total damages for a plaintiff are calculated by combining damages per share and the plaintiff's trading records (with the number of shares bought and sold by the plaintiff on relevant dates). In securities class actions, total damages are typically not covered in expert reports submitted to the court, as class members' trading records only become available after damages claims are filed, typically after a settlement occurs or if a verdict for plaintiffs is reached.

96. *See, e.g.*, Robbins v. Koger Props., Inc., 116 F.3d 1441, 1447 n.5 (11th Cir. 1997).

97. Per-share recoverable damages in Rule 10b-5 cases are the lesser of (1) "out-of-pocket" damages, (2) the losses caused by the alleged misrepresentations (discussed further below), and (3) a cap prescribed by the Private Securities Litigation Reform Act of 1995 (the "90-day bounce-back" provision).

98. *See, e.g.*, Ludlow v. BP, P.L.C., 800 F.3d 674, 682 (5th Cir. 2015) ("At the same time, the 'out-of-pocket measure,' sometimes called the 'price inflation' metric, is often used. Under this theory, 'a purchaser of securities may recover against a defendant . . . only the "difference between the price paid and the '[true]' value of the security . . . at the time of the initial purchase by the defrauded buyer."'").

damages are limited to the losses, if any, caused by the misrepresentations.[99] In the context of damages, losses caused refers to stock price declines, if any, that occurred in the actual world and are attributable to the arrival of information (corrective information) into the market that is causally linked to the misrepresentations.

Quantification of the losses, if any, caused by the misrepresentations requires, as a first step, identification of instances when new corrective information was revealed to the market. The next step is to use tools and techniques from financial economics to measure, for each of those instances, the impact, if any, of such information on stock price. Such measurement is typically conducted using tools and techniques from financial economics. For example, an event study model, discussed below, can be used to estimate the impact that market or industry factors—which are typically not causally linked to the misrepresentations—had on the stock price. In addition, any effect of company-specific information that is not corrective information (confounding information) must be separated from any effect of company-specific information that is. In sum, any losses that resulted from "changed economic circumstances, changed investor expectations, new industry-specific or firm-specific facts, conditions, or other events,"[100] and thus would have occurred even in the absence of corrective information, must be isolated and removed in determining the amount of any stock price declines that are attributable to the misrepresentations.

99. In the context of Rule 10b-5 cases, in *Dura Pharms., Inc., v. Broudo*, 544 U.S. 336, 342–44 (2005), the Supreme Court referred to losses caused as a limitation on out-of-pocket damages: "If the purchaser sells later after the truth makes its way into the marketplace, an initially inflated purchase price might mean a later loss. But that is far from inevitably so. When the purchaser subsequently resells such shares, even at a lower price, that lower price may reflect, not the earlier misrepresentation, but changed economic circumstances, changed investor expectations, new industry-specific or firm-specific facts, conditions, or other events, which taken separately or together account for some or all of that lower price. . . . Indeed, the Restatement of Torts, in setting forth the judicial consensus, says that person who 'misrepresents the financial condition of a corporation in order to sell its stock' becomes liable to a relying purchaser 'for the loss' the purchaser sustains 'when the facts . . . become generally known' and 'as a result' share value 'depreciate[s].'" In the context of Section 11 and Section 12(a)(2), defendants can remove from the damages amount prescribed by statute (*statutory damages*) the portion that did not result from the misrepresentations. *See* 15 U.S.C. § 77(k) regarding Section 11 ("[I]f the defendant proves that any portion or all of such [statutory damages] represents other than the depreciation in value of such security resulting from [the alleged material misrepresentations], such portion of or all such damages shall not be recoverable") and 15 U.S.C. § 77l regarding Section 12 ("if the [defendants prove] that any portion or all of [statutory damages] represents other than the depreciation in value of the subject security resulting from [the alleged material misrepresentations], then such portion or amount . . . shall not be recoverable.").

100. *Dura*, 544 U.S. at 343.

Key Areas of Disagreement in the Economic Analysis of Inflation and Losses Caused

Event study model

Damages analyses in securities cases almost invariably involve an event study model, which is a widely used and generally accepted methodology to analyze changes in stock prices following the revelation of information to the market.[101] Using an event study model to analyze changes in stock price following the release of allegedly corrective information is a common step in the analysis of losses caused.[102]

An event study model can be used to estimate the portion of a stock's return (the percentage change in stock price, including dividends) over some period (the event window) that can be attributed to market and industry factors. The remaining portion, often labeled the residual or abnormal return, may be attributed to company-specific information (i.e., information that affects the company but not the overall market or its industry peers),[103] including any corrective information disclosed over the event window. An event study model can be used to test whether a residual return is statistically unusual relative to the typical level of stock-price volatility (i.e., "statistically significant").[104] Economists interpret residual returns that are statistically significant as evidence that company-specific information caused a stock-price change. Residual returns that are not statistically significant are interpreted as having no cause that can be distinguished from randomness.

There are several common areas of disagreement among experts when using an event study model. Each of these areas illustrates the importance of carefully

101. *See, e.g.*, A. Craig MacKinlay, *Event Studies in Economics and Finance*, 35 J. Econ. Literature 13 (1997), http://www.jstor.org/stable/2729691.

102. An event study model can also be used to analyze changes in stock price following allegedly false or misleading statements, which may be relevant for the analysis of inflation.

103. Note that company-specific information does not necessarily coincide with information disclosed by the company. Sometimes, company disclosures may convey information that moves the overall market and the stock prices of its industry peers. *See, e.g.*, George Foster, *Intra-Industry Information Transfers Associated with Earnings Releases*, 3 J. Acct. & Econ. 201 (1981), https://doi.org/10.1016/0165-4101(81)90003-3; Andrew J. Patton & Michela Verardo, *Does Beta Move with News? Firm-Specific Information Flows and Learning About Profitability*, 25 Rev. Fin. Stud. 2789–2839 (2012), https://doi.org/10.1093/rfs/hhs073. And sometimes, third-party disclosures may convey company-specific information (e.g., a short-seller report analyzing a company's prospects, as discussed in Alexander Ljungqvist & Wenlan Qian, *How Constraining Are Limits to Arbitrage?*, 29 Rev. Fin. Stud. 1975, 2012–14 (2016), https://doi.org/10.1093/rfs/hhw028).

104. The most commonly used threshold for the assessment of statistical significance in event studies is the 5% level. As discussed in David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual, "In practice, statistical analysts typically use levels of 5% and 1%. The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure."

tailoring the event study model in a particular case to the facts and circumstances of that case. Absent such tailoring, experts may reach erroneous conclusions.

First, experts may disagree on whether the event study model adequately captures the effect of market and industry factors. There may be disagreement about whether appropriate market and industry factors[105] have been selected,[106] or about whether the model adequately captures the sensitivity of company stock returns to market and industry factors.[107] A model that overstates or understates the effect of market or industry factors will incorrectly estimate the residual return and thus may lead to an incorrect finding of statistical significance (or lack thereof).[108]

Second, experts may disagree on whether a model adequately estimates the typical level of stock-price volatility used to assess the statistical significance of residual returns.[109] The typical level of volatility may vary significantly over time for a given company stock, as circumstances change.[110] A model that overstates or understates volatility may lead to an incorrect finding of statistical significance (or lack thereof).[111]

Third, experts may disagree on the appropriate event window to analyze residual returns following a particular information release. In an efficient market, a company's stock price will at all times quickly and fully reflect all publicly

105. Market and industry factors are usually modeled using the returns on market-wide and industry indices, respectively. In choosing an appropriate industry index, experts typically balance between an index with more constituent company stocks (less prone to the effect of outlier returns but also potentially less comparable to the stock of interest) and an index with fewer constituents (more comparable to the stock of interest but also more prone to the effect of outlier returns).

106. *See, e.g.,* Exhibit 138 to Defendants' Motion for Summary Judgment, Strathclyde Pension Fund v. Bank OZK, No. 4:18-cv-793-DPM, ECF No. 158–9, (E. D. Ark. Feb. 5, 2022).

107. See, for example, the discussion of the dispute about whether a model adequately captured the sensitivity to industry factors in *Smilovits v. First Solar Inc.*, No. CV12-0555-PHX-DGC, 2019 WL 7282026, at 11–12 (D. Ariz. Dec. 27, 2019).

Academic literature in financial economics shows that the sensitivity of company stock returns to market and industry factors varies depending on the circumstances. *See, e.g.*, Patton & Verardo, *supra* note 103, at 2789.

108. The statistical significance of residual returns is assessed using a t-statistic, which is calculated as the ratio of the residual return to the estimate of typical volatility. Thus a misestimated residual return will lead to a misestimated t-statistic.

109. *See, e.g.,* Defendants' Motion for Partial Summary Judgment, Evanston Police Pension Fund v. McKesson Corp., No. 3:18-cv-6525-CRB, ECF No. 166, (N.D. Cal. June 7, 2021).

110. *See, e.g.*, Robert Engle, *GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics*, 15 J. Econ. Persps. 157, 158 (2001), https://doi.org/10.1257/jep.15.4.157 ("Even a cursory look at financial data suggests that some time periods are riskier than others; that is, the expected value of the magnitude of error terms at some times is greater than at others.").

111. Misestimated volatility will cause a misestimated t-statistic, which is used to assess statistical significance, as discussed *supra* note 108.

available information.[112] Thus, if the stock trades in an efficient market, experts will typically analyze stock-price changes over short event windows (such as the trading day that immediately follows the release). However, under certain circumstances, experts may choose to apply different event windows. For example, in the presence of multiple pieces of information disclosed at different times during the same trading day, event windows shorter than a trading day may be helpful in analyzing the effect of a particular piece of information.

Analysis of "confounding" information

Because an event study model estimates a single residual return for a particular event window, it does not disaggregate the impact of separate pieces of company-specific information arriving within the same event window. This feature of event study models limits its usefulness in securities cases when the arrival of corrective information coincides with the arrival of confounding information.

To value the effect of corrective information when there is confounding information, experts may need to apply available tools and techniques from financial economics (other than an event study model) that can be used to value specific pieces of information (e.g., discounted cash-flow models, valuation based on multiples, earnings response models).[113] To evaluate whether a specific application of such tools and techniques is economically sound, it is necessary to scrutinize the underlying assumptions and determine if they are sufficiently tethered to the allegations, facts, and circumstances of the case. Without such scrutiny, it is not possible to determine a priori whether a particular method or group of methods will be capable of estimating losses caused by misrepresentations consistent with the plaintiff's theory of liability in the case.

Measuring inflation at the time of purchase

Because inflation at each point in time (e.g., at the time of the plaintiff's purchase) cannot be observed from market prices (indeed, it is based on a counterfactual, or

112. This is known as the semi-strong form of market efficiency. *See* Stephen A. Ross, Randolph Westerfield & Bradford D. Jordan, Corporate Finance 346 (6th ed. 2002) ("[S]emistrong-form efficiency requires not only that the market be efficient with respect to historical price information, but that all of the information available to the public be reflected in price.").

113. *See, e.g.*, Richard A. Brealey, Stewart C. Myers & Franklin Allen, Principles of Corporate Finance (11th ed. 2014).

The concept of market efficiency, discussed above, is relevant to the analysis of information that may have impacted the stock price. In an efficient market, only new information can impact the stock price and no price declines can be attributed to the repetition of information that is already publicly available.

but-for, disclosure), it must be estimated. Experts often disagree on how to measure inflation, including whether stock-price declines following the revelation of corrective information can be used to measure inflation earlier in time.

Stock-price declines following corrective information can measure inflation earlier in time only if (1) any price impact of confounding information has been removed, as discussed above, (2) the corrective information matches the but-for disclosure, and (3) the information environment (i.e., market, industry, and company conditions) earlier in time is such that if the company had made the but-for disclosure then, the stock price would have declined by the same amount as it did upon the release of corrective information in the actual world.[114] In its *Goldman Sachs* ruling, the Supreme Court noted that the "inference—that the back-end price drop equals front-end inflation—starts to break down when there is a mismatch between the contents of the misrepresentation and the corrective disclosure."[115] If the facts and circumstances of the case do not support using stock-price declines following corrective information to measure inflation earlier in time, then additional analysis, including analysis based on valuation tools and techniques from financial economics, is typically required.[116]

Estimation of Economic Damages in Antitrust Cases

This section provides examples to illustrate the application of the principles for estimation of economic damages to antitrust cases. The scope of the discussion is limited; a full discussion of damages methods across the complete spectrum of antitrust offenses is beyond the scope of this reference guide.[117]

Threshold Issues

A model for economic damages arising from an antitrust violation must satisfy the causation requirement articulated in *Comcast*, by measuring the economic damages arising only from defendant conduct found to be unlawful.[118]

114. *See, e.g.*, Bradford Cornell & R. Gregory Morgan, *Using Finance Theory to Measure Damages in Fraud on the Market Cases*, 37 UCLA L. Rev. 883, 889–97 (1990).

115. Goldman Sachs Grp., Inc. v. Arkansas Tchr. Ret. Sys., 594 U.S. 113, 123 (2021).

116. As is the case when valuing the effect of corrective information when there is confounding information, there is no a priori reason that a particular method or particular group of methods will be capable of estimating inflation consistent with the plaintiff's theory of liability in a specific securities case.

117. For a comprehensive discussion of antitrust damages, see A.B.A., *Proving Antitrust Damages: Legal and Economic Issues* (2017), and Daniel L. Rubinfeld, *Antitrust Damages*, *in* Research Handbook on the Economics of Antitrust Law 378–93 (Einer Elhague ed., 2012).

118. In *Comcast*, the court held that the plaintiffs failed to present a valid methodology for measurement of class-wide damages because their damages model included the effects of conduct

Further, a model estimating damages arising from an antitrust violation must measure what courts have described as antitrust injury. Plaintiffs must show not only that they suffered economic injury, but additionally that the injury they sustained is "of the type the antitrust laws were intended to prevent and that flows from that which makes defendants' acts unlawful."[119]

In many instances, meeting this requirement can be straightforward. For example, any alleged overcharge that consumers pay due to a price-fixing cartel flows directly from the cartel's subversion of price competition, which the antitrust laws seek to prevent. Hence, damages based on a reliable overcharge computation reflect antitrust injury.

However, not all harms constitute antitrust injury. For example, a firm may be harmed by two rivals that merge to form a single, more efficient competitor, but the harm that firm sustains—an erosion in its market position because it faces a new, more efficient competitor—is the result of increased competition rather than harm to competition.[120] An increase in competition is not something the antitrust laws seek to prevent—quite the opposite, in fact—and as such it is not antitrust injury.[121]

Quantifying Damages in an Antitrust Matter

The foundation of an antitrust damages model is a but-for world that reflects all relevant conditions in the actual world other than the anticompetitive conduct at issue, as explained in the section titled "The Standard General Approach to Quantification of Economic Damages" above.

Models used to compute antitrust damages generally fall into one of two categories, overcharge damages and lost-profit damages. As explained below, the choice of model depends on the nature of the plaintiff's claim and the economic relationship between the plaintiff and defendant.

Overcharge Damages

An overcharge model attempts to measure damages sustained by one of the parties to a vertical buyer–seller relationship by deriving an estimate of what the price in the market would be if the conduct did not occur.

under multiple theories of liability, only one of which was found to be triable on a class-wide basis. *See* Comcast Corp. v. Behrend, 569 U.S. 27 (2013).

119. *See* Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc., 429 U.S. 477 (1977).

120. Cargill, Inc. v. Monfort of Colorado, Inc., 479 U.S. 104 (1986).

121. *See* A.B.A., *supra* note 117, at 19–33. *See also* Jonathan M. Jacobson & Tracy Greer, *Twenty-One Years of Antitrust Injury: Down the Alley with* Brunswick v. Pueblo Bowl-O-Mat, 66 Antitrust L.J. 273 (1998).

For example, a plaintiff that purchased a product from a supplier engaged in anticompetitive conduct might compute overcharge damages as the difference between the supra-competitive price actually paid and the lower price that would have prevailed absent the conduct at issue (i.e., the but-for price), multiplied by the quantity actually purchased. Similarly, a supplier that sold a product to a buyer engaged in anticompetitive conduct (e.g., a monopsonistic conspiracy) may compute overcharge damages as the difference between the higher price it would have received absent the buyer's anticompetitive conduct and the artificially suppressed price actually received, multiplied by the quantity actually sold. Antitrust claims that lend themselves to an overcharge model of damages include horizontal price-fixing, bid-rigging, illegal-monopolization, and exclusionary-conduct claims when brought by a customer or supplier.

Damages experts generally rely on one of two approaches to estimate overcharge damages:

1. A before-and-after model, which compares price during the period in which the anticompetitive conduct is believed to have had an effect (the impact period) with price in a period before (or after) the anticompetitive conduct occurred (the control period).
2. A difference-in-differences model, which compares the change in price from the control period to the impact period with the change in price during the same periods in a reasonably comparable market.

Conceptually, both methods seek to identify price levels absent the impact of the challenged conduct and then estimate the competitive price level in the but-for world. Below is a hypothetical example of a price-fixing cartel to illustrate the application of these methods.

Before-and-after model of overcharge damages

Suppose that the plaintiffs allege harm due to supra-competitive prices charged by a cartel and estimate damages as follows: (1) estimate a regression model with observations on prices over a period of months or years prior to the date at which the cartel took effect,[122] under the assumption that the model describes prices that would have prevailed but for the cartel; (2) use the estimated model to

122. Multiple regression is a statistical method that can be used to measure the relationship between multiple explanatory variables and a variable of interest (e.g., price), and thus distinguish between the effect of defendant's anticompetitive conduct and the effect of other economic factors. A detailed discussion of the use of regression analysis (including to estimate economic damages) is beyond the scope of this reference guide. *See* Daniel L. Rubinfeld & David Card, *Reference Guide on Multiple Regression and Advanced Statistical Models*, in this manual. *See also* A.B.A., *supra* note 117, at 123.

predict the prices that would have prevailed during the damages period but for the cartel; and (3) estimate overcharge damages as the difference between observed prices during the damages period and prices predicted by the regression model, multiplied by the quantity sold.[123]

The validity of such an estimate depends on the validity of the underlying regression model, which must account for the impact on prices of economic factors, other than the challenged conduct, that influence demand and costs. A regression that omits such factors will generally yield an unreliable estimate of damages. The direction of the error depends on the nature of the omitted factors. Two examples illustrate the point:

Example: Negative news during the cartel period about the health effects of the product at issue causes demand to fall, and absent the cartel, the price of the product at issue would fall as a result. A regression model that fails to account for the price effect of the negative news would likely overstate but-for prices during the cartel period and understate damages as a result.

Example: The structure of the market changed due to a merger between rival manufacturers just before the cartel began operating. The price effects of the merger, if any, would persist in the but-for world, as the cartel did not cause the merger. A regression model that failed to separately account for the price effects of the merger would improperly attribute the effects of the merger to the cartel and overstate damages as a result.

A regression model that supports a before-and-after estimate of damages may also yield an unreliable estimate of but-for prices if the conduct at issue also affects the explanatory factors included in the regression. Examining the impact of the conduct at issue on explanatory factors in a regression used to estimate but-for prices is thus a critical step in evaluating a benchmark model of damages.

123. Note that this approach assumes that the impact of explanatory variables is identical during the control and cartel periods. Rather than forecasting but-for prices in the cartel period based on an earlier control period, a damages expert may instead rely on the so-called dummy variable method: (1) estimate a regression model with observations on price using a sample that includes the damages period in addition to the period beforehand; (2) include in the regression specification a dummy variable that measures the difference between prices in the control and cartel periods, holding all other factors constant; and (3) use the point estimate for the dummy variable as the basis for computing overcharge damages. The dummy variable method allows a more flexible specification in which the effects of explanatory variables can differ during the control and cartel periods. *See* Justin McCrary & Daniel L. Rubinfeld, *Measuring Benchmark Damages in Antitrust Litigation*, 3 J. Econ. Methods, no. 1, 2013, at 63–74, https://doi.org/10.1515/jem-2013-0006. *See also* John K. Wald, *Antitrust Damages in Financial Markets*, 16 J. Competition L. & Econ., no. 1, 2020, at 63–73, https://doi.org/10.1093/joclec/nhaa003, for an example of the application of this method in the NASDAQ odd-eighths litigation.

Example: Suppose that the price of the product at issue depends on the price of a particular input, and the regression model includes the price of that input. If cartel participants account for a substantial fraction of demand for the input, then in the but-for world, the price of the input will likely be lower.[124] Thus a regression that relies on the actual input price, which reflects the impact of the cartel, will likely yield a biased estimate of the but-for price of the product at issue.[125]

Difference-in-differences model of overcharge damages

The preceding examples highlight the errors that may arise if a regression-based estimate of overcharge damages omits economic factors that affect price. The omitted factors discussed above—new information and changes in market structure—are observable. In such cases, correcting the bias may be as simple as including appropriate explanatory variables in the regression to measure the impact of the omitted factors on price.

Many cases, however, are not so simple. The omitted factors that affect prices may be unobservable, a variable that can serve as a proxy for unobserved factors may not be available, or the nature of the omitted factor is such that its impact cannot be separately identified using multiple regression. In such cases, damages experts may turn to a difference-in-differences regression model in order to account for the impact of omitted factors.

Rather than estimating the overcharge by examining price before and during the cartel period, a difference-in-differences model isolates the impact of the cartel through comparison to a second, comparable market unaffected by the conduct at issue. A difference-in-differences regression measures the impact of the cartel as the difference between (1) the change in prices in the market of interest, and (2) the change in prices over the same period in a second, comparable market unaffected by the conduct at issue.[126] The key assumption here that identifies the impact of the cartel—the so-called parallel trend assumption—is that prices in the two markets share a common, parallel

124. The cartel suppresses output of the final good and thus suppresses demand for the input in question. If cartel participants account for a substantial fraction of demand for the input, increased demand for the input in the but-for world means that the price of the input will likely increase as well.

125. For a more detailed discussion of this point, see Halbert White, Robert Marshall & Pauline Kennedy, *The Measurement of Economic Damages in Antitrust Civil Litigation*, 6 A.B.A. Econ. Comm. Newsl., no. 1, 2006, at 18.

126. *See, e.g.*, Joshua D. Angrist & Jörn-Steffen Pischke, Mostly Harmless Econometrics: An Empiricist's Companion (2008).

trend, so that absent the cartel, change in prices in the two markets over time would be identical.[127]

For example, suppose the plaintiffs claim that suppliers of a product in a particular geographic region operated a cartel. A damages expert might adopt as a comparator a second, separate geographic market for the same product in which the alleged conspirators do not operate, and use a difference-in-differences regression to account for the effects of unobserved factors and isolate the price effects of the cartel.[128]

Lost-Profits Damages

A lost-profits model is frequently used to measure damages in matters in which the plaintiff and defendant are competitors rather than parties to a vertical, buyer–seller relationship, such as matters involving exclusionary conduct.

While overcharge damages require an estimate of but-for prices, a damages model that seeks to measure lost profits requires a more extensive counterfactual that specifies prices, unit sales (or market share), and costs in the but-for world.[129] Assuming that the plaintiff's fixed costs would not change in the but-for world, lost-profit damages for a firm harmed by a rival's anticompetitive conduct are given by the following formula[130]:

$$D = (P_{but-for} - C_{but-for}) \times Q_{but-for} - (P_{actual} - C_{actual}) \times Q_{actual}$$

A model that estimates lost profits must carefully consider the effects of the conduct at issue on each element of the plaintiff's profits in the but-for world.

Price Effects. Because conduct that partially or completely excludes a firm from a market reduces competition and allows the defendant to increase prices, prices in the but-for world will likely be lower than observed prices. Indeed, the prospect of increasing prices creates an incentive for firms to engage in exclusionary conduct.[131] That said, in some circumstances plaintiffs may assert that in the but-for world, they would be able to charge higher prices, as the

127. Note that the regression specification should still include all observable factors that affect price and are independent of the cartel (i.e., do not vary as a result of the cartel), just as with the before-and-after model discussed above.

128. *See, e.g.*, R. Forrest McCluer & Martha A. Starr, *Using Difference in Differences to Estimate Damages in Healthcare Antitrust: A Case Study of Marshfield Clinic*, 20 Int'l J. Econ. Bus. 447 (2013), https://doi.org/10.1080/13571516.2013.800323.

129. Some damages experts may specify damages using a model of but-for profit margins rather than modeling underlying prices and costs.

130. Costs shown here should be read as average cost per unit of output, not as variable cost.

131. Note, however, that some theories of liability, such as predatory pricing, imply higher prices in the but-for world.

challenged conduct prevented them from effectively competing and forced them to offer prices lower than those they otherwise would have offered.[132]

Quantity Effects. Economic theory teaches that if prices increase, output declines. Thus, if the effect of the conduct at issue is to raise prices, lower prices in the but-for world will generally be accompanied by higher output. However, the critical question is not output for the market as a whole but whether the plaintiff's output would increase absent the challenged conduct.

Lower prices and higher output have offsetting effects on profits: All else being equal, lower prices reduce profits, but higher output increases profits. The net impact of these effects is an empirical question whose answer depends on the elasticity of demand for the products or services at issue. If demand is elastic, quantity effects will outweigh price effects, profits in the but-for world will increase, and all else being equal, the plaintiff's profits in the but-for world would be higher. If demand is inelastic, profits in the but-for world will decrease, and all else being equal, the plaintiff's profits in the but-for world would be lower.

Cost Effects. The impact of the challenged conduct on the plaintiff's costs is likewise an empirical question; the plaintiff's costs can increase or decrease as a result of the challenged conduct. In some cases, the conduct at issue may raise the plaintiff's costs. Indeed, the specific objective of exclusionary conduct is often raising a rival firm's costs. Additionally, rather than directly raising input costs, the conduct at issue may increase the plaintiff's costs by reducing the plaintiff's output and thereby depriving the plaintiff of the benefit of economies of scale. In such cases, the plaintiff's unit costs would decline as output increases in the but-for world.

On the other hand, the plaintiff's costs in the but-for world may be higher if the output increase contemplated in the damages model exceeds the plaintiff's actual capacity to produce, market, or distribute the products or services at issue and would require additional investment. In such cases, the plaintiff must establish that it had the wherewithal to increase its capacity in the but-for world, and estimated damages should reflect the costs of any incremental investments necessary to expand the plaintiff's capacity.

Models of lost profits frequently rely on the assumption that the period prior to the conduct at issue provides a reliable basis for predicting market conditions in the but-for world. As explained in the section titled "Considering All the Differences in the Plaintiff's Situation in the But-For Scenario Versus Assuming That Many Aspects Would Be the Same as in Actuality" above, this assumption requires careful scrutiny, as it may yield an unreliable estimate of damages. The purpose of the defendant's challenged conduct may have been to forestall

132. *See, e.g.*, ZF Meritor, LLC v. Eaton Corp., 696 F.3d 254 (3d Cir. 2012). ZF Meritor's lost-profits damages model relied on projections that assumed that the exclusionary loyalty discounts offered by defendant Eaton forced ZF Meritor to offer deep discounts that it otherwise would not have offered.

changes in competitive conditions (e.g., by deterring entry) and maintain the status quo, in which case market conditions in the period prior to the challenged conduct are not a reliable basis for predicting market conditions in the but-for world. Suppose, for example, that the market for a particular product was a duopoly, and that one firm (the defendant) entered an exclusive dealing agreement with a downstream purchaser that foreclosed the second firm (the plaintiff) and deterred the entry of new competitors. Then a damages model that assumes a duopoly in the but-for world would yield an unreliable conclusion, as it fails to account for the competitive impact of nascent competitors that would have entered the market but for the challenged conduct.

Estimation of Economic Damages from Loss of Personal Income

Many of the disputes that arise in estimating damages for lost personal income can be resolved by carefully applying the standard damages approach. Damages are the difference between the but-for and actual worlds, where the actual world reflects any mitigating factors. Estimating such damages can also involve issues that are unique, such as estimating losses over a person's lifetime, valuing fringe benefits, estimating lost income in wrongful death cases, or estimating compensation for shortened life expectancy. We discuss these issues below.

Estimation of Losses Over a Person's Lifetime

In nearly all cases involving lost income, the effects continue past trial and sometimes until the plaintiff's death. Therefore, quantifying damages for loss of personal income necessarily involves projecting the plaintiff's work history and retirement. Conceptually, the estimate of income for each year, either but-for or actual, is the expected income multiplied by the probability that the person will be working for that year. The probability that the person will be working for that year is the product of the probability that the person will survive the year, the probability that the person will be in the labor force, and the probability that the person will be employed, given that the person worked in the prior year.[133] We refer to this as the standard method for estimating personal losses.

In many cases, such as those involving wrongful termination, the projection that the plaintiff was working is the same for both the but-for world and the actual

133. Except for wrongful-death cases, the probability that the persons will be working is usually 1 for both the but-for and actual cases. In wrongful-death cases, the probability is still usually 1 in the but-for case but 0 in the actual case.

world. However, in wrongful death cases and some personal injury cases, these projections may differ,[134] and the expert will need to compute separate projections for the but-for and actual worlds before taking the difference between the two.

These projections may rely on data from government agencies, including the Bureau of Labor Statistics (BLS) and the Social Security Administration (SSA).[135] These data include tables on survival, workforce participation, and employment. The expert usually needs to manipulate this information in order to generate the conditional probabilities needed.

In order to simplify the calculations, the expert may use the person's expected lifespan and retirement age based on the person's age at trial using standard tables from the SSA and BLS. Then the expert need only sum the discounted losses for each year until the expected age at death. This method, often referred to as the life-expectancy method, will considerably simplify the calculations associated with estimating lost retirement benefits. However, the standard and the life-expectancy methods will usually generate the same estimate of losses only if the expert is assuming that the expected discounted income in each year is the same—that is, that the expected increase in income is offset by the discount rate (see section titled "Disagreements About the Discount Rate to Value the Impact of the Alleged Harmful Act in Periods After the Valuation Date" above).

Calculation of Fringe Benefits

Fringe benefits are often a component of lost pay and may include medical insurance and retirement benefits such as Social Security. Although sick days and vacation are also fringe benefits, they are included already in lost-pay calculations. An exception occurs if these days are accrued but not taken, where, for example, the employee lost the cash payout that would have occurred at a normal termination or lost the benefit of future days off with pay.

Medical insurance benefits

In the following discussion, we assume that the plaintiff no longer has the benefit of the employer-provided insurance as a result of the actions of the defendant. Such situations typically arise in wrongful-termination or wrongful-death cases.

134. This situation may arise in a personal injury case if the injured person is less likely to be able to work as a result of the accident. If so, then the person may have an increased likelihood of leaving the labor force for each age compared to the likelihood prior to the incident. Similarly, the injured person may be more likely to retire at an earlier age.

135. Data from additional providers may be available with a finer level of granularity than SSA or BLS data (e.g., with respect to location or occupation). However, the reliability of these data may be disputed because of questions about the methodology used to generate the tables.

Calculating damages for lost medical insurance is straightforward if the plaintiff can purchase insurance under COBRA, or from their current employer, or on the open market. Then the value of the lost medical insurance is the employee's portion of the premium. If insurance coverage available to the plaintiff differs significantly between the but-for and the actual worlds, then the expert will need to project the impact of the difference in policy coverage.

If the plaintiff chooses not to purchase insurance even though the option is available, or the plaintiff is unable to purchase insurance,[136] then the plaintiff may argue that the value of insurance is the sum of their actual expenses less the premium in the but-for world. The defendant will likely respond that the plaintiff assumed the risk that the incurred medical expenses could exceed the plaintiff's portion of the premium, and therefore that the defendant's responsibility should be limited to the plaintiff's portion of the premium forgone. The plaintiff may counter that their pay was insufficient to afford the insurance.

Retirement benefits

For lost retirement benefits, the issues are similar to those involving lost medical benefits, but the calculations are more complex. There are basically two types of retirement plans: defined benefit plans and defined contribution plans. Defined benefit plans are those where the benefits paid out after retirement are guaranteed to be a definite amount upon retirement. In contrast, defined contribution plans are those where the employer makes a predefined contribution for the employee but the benefits paid out depend on the return earned on the money invested.

The expert can calculate both types of retirement plans on the basis of either the amounts paid in or the amounts paid out by the employer. If the expert uses the amount the employer paid for the benefits, which may be a function of the amount the plaintiff earned, then the calculation is analogous to computing the loss in plaintiff's earnings. The disadvantage of this approach is that the amounts paid in may not adequately predict the benefits paid out, particularly when the plan is a defined benefit plan. We discuss this topic below in connection with Social Security benefits, where the problem is particularly acute.

Defined benefit plan

To determine the present value of the benefits received under a defined benefit plan, the calculation is simplified if the expert uses the life-expectancy method

136. For example, a preexisting condition may make it difficult to purchase insurance on the open market or may limit the plaintiff's coverage to exclude a preexisting condition either permanently or for a period of time.

to calculate the plaintiff's losses. In this situation, the expert must determine the number of years that the plaintiff would have worked at the firm upon retirement, and the plaintiff's retirement age, expected lifespan, and salary at the firm over time. These factors must be consistent with the expert's assessment of the projected trajectory of the plaintiff's employment in both the but-for and the actual worlds.

However, if the expert instead uses probability tables for each year, then the calculation is more complex. For each year where the plaintiff may cease to be in the labor force for reasons other than death, the expert must determine the likelihood that the plaintiff is receiving benefits from the plan because of disability (if the plan permits) or retirement. Complicating this determination is that the payout from the plan may depend on the age at retirement. Thus, the calculation must incorporate the probability for each possible payout. Depending on the plan, defining possible outcomes can be extremely complex.

A common example of a defined benefit plan is Social Security.[137] Determining benefits from Social Security can be forbiddingly complex because the number of potential outcomes is so large. For example, a person can retire at almost any age, and disability payments are made if the person is unable to work. In addition, calculating the benefit at any age depends on the person's average salary over the most recent thirty-five years. If Social Security benefits are critical to the magnitude of damages, the expert may choose to simplify the calculation by relying on the life-expectancy method.

Defined contribution plan

For a defined contribution plan, the expert's task is to project the employer's contribution, the number of years that the employee would have worked at the firm, and the employee's age at retirement. Generally, this determination is straightforward because it is based on the same factors the expert would use to project the employee's salary in the but-for and actual worlds. The present value of the employer's contributions will be the expected payouts from the plan.

Wrongful Death

Generally, calculation of economic damages for wrongful death depends on whether the claimant is a relative of the decedent or is the estate. If the claimant is a relative of the decedent, economic damages are limited to the economic value that the relative would have received had the decedent lived. If the relative

137. Social Security is generally regarded as a defined benefit plan although it has some elements of a defined contribution plan.

is a dependent, the recovery may be substantial, whereas if the decedent was a child or unmarried and childless and the only relatives are parents, the recovery may be small, because most parents receive little economic value from their children. In contrast, if the claimant is the estate, the recovery may include all of the lost economic value.

Where the claimant is a relative, damages from lost wages may be reduced to reflect the decedent's own consumption spending had the person continued living. Such expenses may be relatively small if the claimant is a spouse with children under the theory that much of the decedent's income would have been spent to support the dependents. If decedents have no children or if there is another earner in the family, offsets for the spending of decedents on themselves may be higher.

Shortened Life Expectancy

An important issue is whether a plaintiff may recover compensation for shortened life expectancy caused by an injury. This issue may arise, for example, in medical malpractice cases in which a doctor fails to diagnose and treat a condition or where a surgeon fails to remove a medical device used during surgery.

A related issue is whether dependents in a wrongful-death action may recover economic damages for support the decedent would have provided had the decedent lived—that is, whether such damages can be recovered over the remainder of the decedent's expected lifetime, had he lived. Quantifying such damages requires a projection of the decedent's life expectancy using the methods discussed above.

Acknowledgments

The authors would like to express their gratitude for the substantial contributions of Eduardo Hurtado and Neill Norman in the development of this reference guide.

Glossary of Terms

avoided cost. Cost that the plaintiff did not incur as a result of the harmful act. Usually it is the cost that a business would have incurred in order to make the higher level of sales it would have enjoyed but for the harmful act.

before-and-after damages. A damages model based on the assumption that the period prior to the challenged conduct (or at some point after the challenged conduct) provides a reliable basis for conditions in the but-for world, absent the challenged conduct.

but-for analysis. Description of the plaintiff's economic situation but for the defendant's harmful act. Damages are generally measured as but-for value less actual value received by the plaintiff.

capitalization. Calculation of today's equivalent to a past dollar to reflect the time value of money and risk. If the capitalization rate is r, the capitalization factor applicable to one year in the past is:

$$\text{capitalization factor} = (1 + r).$$

The capitalization for two years is the capitalization factor squared; for three years, it is the capitalization factor cubed, and so on for longer periods.

compound interest. Interest calculation giving effect to interest earned on past interest, as opposed to simple interest, which does not. Simple interest at interest rate r over a number of periods t and on a principal amount P is computed as $P(1 + rt) - P$, while compound interest earned over the same period is computed as $P(1 + r)^t - P$. Thus, for example, given an interest rate of 20%, the compound interest earned on a lost dollar of income three years earlier is computed as $(1 + 0.20)^3 - 1 = \$0.73$, while simple interest is $(1 + 0.20 \times 3) - 1 = \$0.60$.

difference-in-differences. A regression methodology that seeks to identify the impact of the challenged conduct by comparing the change in outcomes (e.g., prices or market share) over time between a group (the treatment group) exposed to the challenged conduct, and a second group (the control group) not exposed to the challenged conduct.

discount rate. Rate used to discount future cash flows.

discounting. Calculation of today's equivalent to a future dollar to reflect the time value of money and risk. If the discount rate is r, the discount factor applicable to one year in the future is:

$$\text{discount factor} = 1/(1 + r).$$

The discount for two years is the discount factor squared; for three years, it is the discount factor cubed, and so on for longer periods.

earnings. Economic value received by the plaintiff. Earnings could be salary and benefits from a job, cash flows from a business, royalties from licensing intellectual property, or the proceeds from a one-time or recurring sale of property. Earnings are measured net of costs. Thus, lost earnings are lost receipts less costs avoided. Note that the term earnings can also be used to mean the accounting concept of net income, which is not a purely economic concept.

fixed cost. Cost that does not change with a change in the amount of products or services sold.

prejudgment interest. Interest on losses occurring before trial.

present value. Value today of money due in the past (with interest) or in the future (with discounting).

price erosion. Effect of the harmful act on the price charged by the plaintiff. When the harmful act is wrongful competition, as in intellectual property infringement, price erosion is one of the ways that the plaintiff's earnings have been harmed.

regression analysis. Statistical technique for inferring stable relationships among quantities. For example, regression analysis may be used to determine how costs typically vary when sales rise or fall.

variable cost. Component of a business's cost that would have been higher if the business had enjoyed higher sales. See also avoided cost.